

Unapproved tests on a chip

Prenatal genetic tests can now check for abnormalities in developing fetuses, but there is still no regulatory framework for them in the United States.

For more than a decade, genetic tests have been on the way that would tell patients about genetic variations that might increase their susceptibility to disease. And all the while, bioethicists have been warning the US government of the impending need to regulate such tests when they arrive. Now hundreds of these tests are available, including powerful prenatal tests that use microarrays to scan hundreds of genes in unborn children. Yet despite the warnings, the tests remain effectively unregulated.

Last December, Affymetrix of Santa Clara, California, became the first company to get approval from the US Food and Drug Administration (FDA) for a microarray chip for use as such a diagnostic device. The chip can check for genetic differences that might predict a patient's tolerance to drug treatment. Now, researchers at Baylor College of Medicine in Houston, Texas, are offering pregnant women a genome-scanning test that checks for abnormalities in developing fetuses (see page 733). But because of the way this test was developed, it is not currently subject to FDA review — and neither are tests made by private companies that can do prenatal screening but haven't yet been used to do so. The FDA has declined to say whether it intends to review these sorts of tests.

One major point of confusion is whether the FDA has the power to regulate tests that are developed within a particular laboratory and are then used only in that lab, without being sold to outsiders. In 1997, the agency said that it would regulate these 'home brew' tests under the same rules that it uses to regulate other medical devices. But in 2003, when it issued further guidance on the ingredients that are used to make home brews, it decided that it didn't have the power to regulate them after all. Instead, the FDA said it would regulate only the 'analyte-specific reagents' — the ingredients, such as antibodies or nucleotide sequences, that make up the test. And it will only regulate the analyte-specific reagents if they are sold to the

maker of the test; if the reagents are developed in-house, neither the reagents nor the test are subject to regulation by the FDA.

Companies that sell microarrays for others to use in genetic tests do have to register with the FDA, and must submit information supporting their ability to make the arrays properly to the regulator for approval. These rules also hold for companies or labs that make key components of genetic tests, such as gene probes. However, private labs that make and use their own microarrays for genetic testing are not subject to FDA scrutiny.

Additionally, some companies are stretching the interpretation of the rules by marketing genetic tests that may be of dubious value (see *Nature* 426, 107; 2003). Because the companies make and use their own testing materials, they don't have to register under the regulations governing companies that sell microarrays to others. The unfortunate end result is a free-for-all in the marketing of genetic tests to the public, as well as widespread confusion among researchers and laboratories about their regulatory obligations.

The introduction of prenatal microarray tests into this picture creates additional urgency for the regulators to act. If misdiagnoses occur, children may be born with an unexpected disease, or fetuses may be terminated on the basis of false information. Let's hope that such episodes don't have to be documented before the FDA acts to clarify its role as a regulator of genetic testing. If the agency finds that it lacks the authority to properly supervise the makers and users of home-brew genetic tests, then Congress should intervene to grant it that authority. ■

"The unfortunate end result is a free-for-all in the marketing of genetic tests to the public, as well as widespread confusion among researchers and laboratories."

Launching a business

There's little evidence that commercial approaches can radically reduce the cost of getting into space.

It has long been an item of faith among some space aficionados in the United States that private enterprise can, if given the chance, sharply reduce the cost of rocket launches. For this vocal group, the government — and NASA in particular — has always been the enemy. What's necessary, they believe, is a commercial launch business unfettered by bureaucratic oversight.

The emergence of such a business has been hampered by its limited range of customers. A few operators of commercial telecommunication satellites aside, the only reliable customers for would-be rocket

makers in the United States are government agencies, primarily NASA and the Department of Defense. In the 1990s, it looked for a while as though the satellite business might rapidly expand, as plans were laid to have swarms of small communications satellites circling the Earth. A number of would-be rocket builders opened offices, hoping to capture some of that business. But when the new satellite businesses failed to materialize, the rocket companies disappeared.

Today, the most prominent player in the private rocket business is Internet tycoon Elon Musk, whose California-based company SpaceX has pledged to bring down the cost of launching materials into orbit by an order of magnitude (see page 736).

Musk starts with several advantages. His rocket, Falcon 1, which is sitting on the Kwajalein atoll in the Pacific awaiting its first launch, was produced by a small design team. It doesn't have the overheads of the large corporations such as Boeing or Lockheed Martin that



build existing US rockets. The rocket project is financed by Musk himself, and if it works as advertised, there's a real chance that SpaceX could offer to launch payloads at a lower cost than existing options.

But that would be only half the battle. The main things constraining the development of new launch options are the low number of customers and the emphasis that these customers place on reliability, as opposed to cost.

Government agencies are by far the largest customers for rocket launches, and they would like to bring costs down. But reliability remains a greater priority. Take a high-value payload such as the \$4.5-billion James Webb Space Telescope planned for launch sometime in the next decade, or the \$500-million New Horizons Pluto probe scheduled to take off from Cape Canaveral on an Atlas V rocket next month. When the satellite costs far more than the rocket ride, the project manager will pay extra to make sure the spacecraft is delivered safely to orbit. A few tens of millions of dollars in savings wouldn't matter much considering the cost of failure. Similar considerations influence operators of telecoms satellites, who can seldom afford to lose them or delay their arrival into space.

The only remaining customer potential lies with space tourism.

But even assuming that a few dozen millionaires visit Earth orbit each year by 2020, the market will remain commercially insignificant. In any case, the space tourists — or at least their insurance companies — may also favour proven reliability over a cheap ticket.

Openings will still arise for the development of more space-launch options on the margins. NASA, for example, is now considering relaxing its traditional insistence on several layers of oversight and inspection for flights that will take food and water to the space station. The agency would instead pay for a delivery service and let the launch provider assume responsibility for the success of the launch.

Such approaches will help to spur on people such as Musk and establish whether they can indeed build a reliable track record in the space-launch business. Until they have done so, the suggestion that entrepreneurial activity can make a substantial difference to the cost of space travel should still be considered pie in the sky. ■

"Even assuming that a few dozen millionaires visit Earth orbit each year by 2020, the market will remain commercially insignificant."

Enough, already

No convincing case has been made for increasing the amount of plutonium held at a Californian lab.

The US Department of Energy is planning to double the amount of plutonium that can be stored at the Lawrence Livermore National Laboratory in California. Under new rules announced last week, the nuclear-weapons lab can keep up to 1,400 kilograms, or enough for around 300 bombs.

Not surprisingly, antinuclear activists are up in arms about having so much bomb-grade metal in such a heavily populated area. But researchers who want the US nuclear-weapons laboratories to set a good example for the rest of the world should be equally dismayed at the plan.

Since 1992, the United States has maintained a moratorium on the testing and development of new nuclear weapons. There's no real need for this research lab, which accommodates an outstanding civilian research programme next to its weapons-related activity, to be playing with this quantity of plutonium.

Livermore is expected to use some of the expanded inventory in nuclear-weapons research, including experiments at the National Ignition Facility (NIF), a massive laser facility that will recreate some of the conditions inside nuclear weapons at detonation. The facility's original function was to perform such experiments on hydrogen isotopes, rather than plutonium. Officials at the Department of Energy never formally excluded the option of using plutonium in the NIF, but a 1995 report prepared by scientists in the department's non-proliferation office warned that its use at the facility could be seen as provocative by other nations.

The other main reason why Livermore wants to hold more plutonium, according to energy-department documents, is that it will start to lay the groundwork for the renewed mass production of

plutonium pits, used in US nuclear weapons. Livermore will be charged with developing new technologies for manufacturing the pits, for use at a proposed industrial-sized production facility. But questions remain over whether this facility is either necessary or appropriate, and this year Congress declined to appropriate the money needed to begin planning for its construction.

Most of Livermore's new plutonium stocks would be shipped there from the Los Alamos National Laboratory in New Mexico, where the Department of Energy's track record in handling plutonium does not inspire much confidence. According to a report released on 29 November by the Institute for Energy and Environmental Research, a watchdog group based near Washington DC, Los Alamos has managed to lose between 300 kg and 600 kg of the material over the years. The group suggests that much of it was dumped indiscriminately in the desert during the early days of the nuclear age, or was mislabelled when shipped off elsewhere for long-term storage.

And Livermore has had its own problems with plutonium. In January, its plutonium facility, where scientists work with the metal under heavily controlled conditions, was shut down amid safety concerns. Problems cited at the time included cracks in the building's ventilation systems and poorly constructed 'hot boxes' for handling the metal. The facility was allowed to reopen at a reduced capacity last month.

In light of all this, Livermore's plan to double its inventory of plutonium is ill-advised. A case for plutonium experiments at the NIF has not been made, even to review groups that have the security clearance needed to assess it. And the laboratory is wasting its time researching pit production for a facility that may never actually be built. For a mixed-use scientific facility in a residential area, 700 kg of plutonium is enough, already. ■

"The laboratory is wasting its time researching pit production for a facility that may never actually be built."

RESEARCH HIGHLIGHTS

Flavour of the week

Science **310**, 1495–1499 (2005)

Mice lacking ATP signalling in their taste buds cannot tell a food pellet from the finest cheese. A new mouse study suggests that ATP tells nerves about basic flavours.

Until now, serotonin was thought to convey this information. But Thomas Finger and Sue Kinnamon of the Rocky Mountain Taste and Smell Center in Aurora, Colorado, and their colleagues found that mice lacking serotonin signalling could respond to basic tastes such as sweet and bitter.

Such responses were deficient in mice lacking receptors thought to mediate ATP signalling. These receptors coat the surface of nerves that innervate taste buds (pictured) and lead to the brain. Taste buds apparently release ATP when exposed to various substances, activating the nerves.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

OMIKRON/SPL

BIOLOGY

Amyloid advantage

PLoS Biol. doi:10.1371/journal.pbio.0040006 (2006)

Plaques of fibrous amyloid protein are associated with a range of human pathologies, including Alzheimer's disease. Although bacteria and fungi are known to use this unusual protein structure, little effort has been expended on finding whether it is a natural feature of mammalian cells.

Jeffery Kelly of The Scripps Research Institute in La Jolla, California, and his colleagues have shown that it is. They used dyes designed for post-mortem identification of amyloid plaques, and recreated plaques *in vitro* to show that a protein involved in making the dark pigment melanin adopts an amyloid structure.

The team suggests that amyloid proteins may exist in other mammalian cell types, and say that their work could have implications for treatments targeted at disrupting pathological amyloid formation.

CELL BIOLOGY

Divide and conquer

Cell **123**, 787–801 (2005)

Cloning of embryonic stem cells by transplanting the nuclei of adult cells into eggs rarely succeeds. But Marcel Méchali and Jean-Marc Lemaître and their colleagues at the Institute for Human Genetics in

Montpellier, France, have found a trick that could make it more successful.

The cells containing the nuclei must first undergo mitosis, or cell division, to prepare them for the rapid cell divisions in early embryonic development. The researchers treated the nuclei of frogs' red blood cells with an extract from mitotic unfertilized frog eggs. Nuclei treated this way could replicate their DNA as quickly as early embryonic cells — once every 30 minutes.

The work identifies mitosis as the crucial step for reprogramming nuclei for development and could, the authors say, open new perspectives in animal cloning.



COMPUTER SCIENCE

Virus buster

Nature Phys. doi:10.1038/nphys177 (2005)

A model of computer networks suggests that companies selling antivirus software should make their cure more like the disease.

At present, computer users gain immunity 'statically' by downloading antivirus updates from a central storage server. But new work indicates that infections might be better limited if antiviral software could spread immunizations through a network, in much the same way as the original virus spreads.

The system could be made even faster if the Internet were sprinkled with special virus-detecting nodes that alerted each other to new infections, write Eran Shir of Tel-Aviv University in Israel and his colleagues. Each of these special nodes could initiate the spread of a 'good' epidemic among the regular nodes.

GENETICS

A purrfect model?

Hum. Mol. Genet. **14**, 3587–3593 (2005)

Cats might not be man's best friend, but they may be set to do us a big favour by helping people who suffer from a life-threatening heart condition. Familial hypertrophic cardiomyopathy is an inherited heart disease that strikes young adults and can lead to sudden death. The disease also afflicts the Maine Coon breed of cats (pictured left).

A. WITZE

Researchers led by Kathryn Meurs of the Ohio State University in Columbus have sequenced the cat equivalents of genes that can cause the human disease if they are spelt incorrectly.

The team now reports that a mutation in the cat equivalent of one of these human genes, known as *MYBPC3*, can also cause the feline form of the disease. The finding suggests that Maine Coons could provide the first useful large animal model for studies of this disease.

METEORITICS

Rocks' clocks reset

Earth Planet. Sci. Lett. doi:10.1016/j.epsl.2005.09.007 (2005)

Planetary geologists have long known that a rare group of meteorites known as SNCs (pictured right) originally came from Mars, and were blasted to Earth after other space rocks hit the martian surface. But one type of SNC, the shergottites, posed a dilemma. Dating of the rocks had suggested they were roughly 180 million years old. However, the red planet shows no widespread signs of having been bombarded that recently.

Now a team led by Audrey Bouvier of the Ecole Normale Supérieure in Lyon, France, may have an answer. The researchers measured lead isotope ratios in four shergottite samples found them to be much older than believed — 4 billion years, an age consistent with the lack of recent cratering on the martian surface.

The previous, isotope-based dating may have given younger ages because acidic groundwater percolated through the rocks in the relatively recent past, resetting their isotopic clocks, the team suggests.

ORGANIC CHEMISTRY

Get shorty

Angew. Chem. Int. Edn **44**, 7549–7553 (2005)

Where molecules are concerned, chemists can be merciless. They have put carbon–carbon single bonds on the rack before, seeing how far they can be stretched. Now Deborah Huntley of Saginaw Valley State University in Michigan and her colleagues apply the thumbscrews, compressing C–C bonds to uncomfortable extremes.

Normal C–C single bonds are about 0.154 nanometres long. Using quantum

chemical methods, the researchers propose molecules that they predict have C–C bonds as short as 0.132 nanometres. This shortening was achieved by putting simple alkanes into molecular cages or by sandwiching the bonds within highly constrained hydrocarbon frameworks.

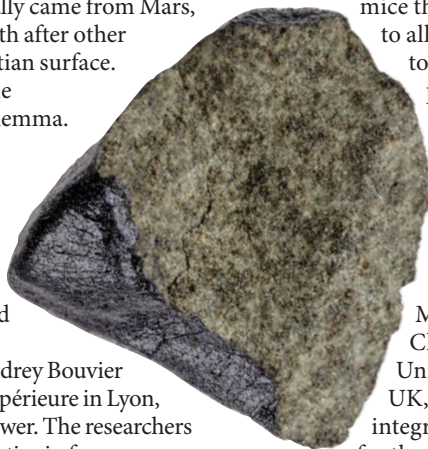
CELL BIOLOGY

Protein shake-up

J. Cell Biol. **171**, 717–728 (2005)

Integrins are proteins that help attach cells to their extracellular matrix. They also help to transmit external signals into cells, and ensure that the mobile molecules involved in cell signalling pathways are positioned in the right place at the right time.

β_1 integrin is known to be crucial for the normal development of embryos. Using mice that were engineered to allow the β_1 integrin gene to be switched off at particular times, researchers now show that this integrin is also required for the normal development of mammary glands in pregnant adult females. The team, headed by Matthew Naylor and Charles Streuli from the University of Manchester, UK, demonstrated that β_1 integrin is, in addition, crucial for the mouse's ability to nurse her young.



NATURAL HISTORY MUSEUM

OPTOELECTRONICS

A little light on the matter

NanoLetters doi:10.1021/nl051811+ (2005)

Miniature organic light-emitting diodes (OLEDs), some as small as 60 nanometres across, could help scientists to make light work of nanoscale tasks. The tiny lights, their inventors say, may prove useful for quantum communication or in photopatterning nanomaterials.

The OLEDs, made by Zakya Kafafi and her colleagues at the Naval Research Laboratory in Washington DC, rely on a light-emitting polymer called MEH–PPV. The polymer is packed inside cylindrical nanoholes etched about 100 nanometres deep into a film of silicon nitride. Each cylinder acts as an independent OLED.

Tests showed that the nanodiode's electrical and light-emitting properties are much like those of a larger reference OLED.

JOURNAL CLUB

H. Eugene Stanley
Boston University,
Massachusetts

A physicist ventures into the no-man's-land of water to find the source of its unusual properties.

Instead of behaving like other liquids, water acts as if there were mathematical singularities in its thermodynamic properties such as compressibility and specific heat. These abnormalities appear at about -45°C , where such functions would become infinite in value.

Water has such wide importance that scientists from many disciplines (including this author) seek a coherent explanation for this unusual behaviour. Indeed, in July, *Science* included water on its list of the 125 most important open questions in science today.

One theory that promises to unify all of water's strange properties is the liquid–liquid (LL) critical-point hypothesis: it says that liquid water possesses a critical point, below which it can switch from one phase, a high-density liquid, to another phase, a low-density liquid. The LL critical-point hypothesis has received a generous amount of theoretical support, but experimental proof has remained elusive because the LL critical point appears in the 'no-man's-land' of temperatures where bulk water is always frozen.

Recently, Sow-Hsin Chen's team at the Massachusetts Institute of Technology in Cambridge succeeded in probing these low temperatures, using the clever trick of confining water to nanopores so narrow that the liquid could not freeze. The researchers discovered a transition between two dynamic behaviours known as fragile and strong, suggestive of the two phases (L. Liu *et al. Phys. Rev. Lett.* **95**, 117802; 2005).

Since then, Chen and members of my group have collaborated to show that the experimental results are best explained by the existence of a critical point (L. Xu *et al. Proc. Natl Acad. Sci. USA* **102**, 16558–16562; 2005) — so, at last, there is clear evidence of the LL critical point.

NEWS

UK animal labs still under siege

Construction work restarted last week on a biomedical research facility at the University of Oxford. Completion of the centre has been delayed for 18 months because of protests by animal-rights groups.

The original building contractor pulled out in July last year because of threats to its workers. Since then, no work has been carried out on the building, which should have been completed last autumn.

The university had always stressed that it would proceed with the facility, which will replace and upgrade most of its animal housing. It has engaged new contractors whose identities are being kept secret — although one animal-rights group says it can and will reveal them. On 30 November workers delivered building material to the site under police protection. Complex and expensive security arrangements have been put in place to protect contractors and the site.

The travails of the £20-million (US\$35-million) project typify the problems that British scientists who use animals in their research have had to face in the past three decades of pressure from extremist animal-rights groups. In the past ten years, activists have started to target individual scientists — this has created a sort of siege mentality.

But official figures show that in the past year or so the number of extremist attacks has fallen. New laws proposed shortly after the halt to building work in Oxford last year have helped. Introduced this summer, they make it illegal to protest outside people's homes if this causes "harassment, alarm or distress", and to use harassment to inflict economic damage on a company. At the same time researchers have changed tactics. Instead of avoiding the public eye, they are being more open about their work and educating the public about the benefits of animal research to medicine.

But an informal survey by *Nature* reveals that the trauma of the past few decades still has effects. The possibility of personal risk puts strain on researchers, their institutions and the relations between the two. Nearly all those who spoke to *Nature* asked for their names not to be published.

One result of continuing tensions is that



Building tension: contractors in Oxford hope that anonymity will allow them to evade protestors.

those entering a scientific career prefer not to work with animals. "Very few people are willing to stake their career on primate science or whole-animal studies," says one primate researcher from a major university. Technicians for animal houses are also hard to recruit, often being put off by the security measures. "Who would be prepared to go ex-directory with their home phone, have their car deregistered with the licensing authority and learn what to do if they are followed?" asks one animal-house manager.

Fearful of the risks to their staff and property, some universities have been careful to keep their animal research low profile. Around ten universities continue to support primate facilities, but one neuroscientist at another big research university says he gave up after becoming the only one using them. "There was no absolute ban," he says, "but it was not encouraged because it was seen as producing a risk to the university."

For the same reason, universities are loath to do building work on animal housing, although

there is no suggestion that animal welfare has been compromised. "The infrastructure here is getting old," says a physiologist using mice at a major university. "We are upgrading, but the nature of that upgrade is determined by needing a low profile."

Researchers also say they are hampered by the UK Home Office's strict regulations on animal research. "I wouldn't do this personally," says a senior neuroscientist at one large research university, "but I am aware of some colleagues who go to mainland Europe, because they feel there are too many hoops to jump through in Britain." He says primate experiments have been taken abroad to avoid Home Office rules that were introduced after the projects had been approved by funders.

Roger Morris, head of biochemistry at King's College London, acknowledges the problems still facing researchers, but says things have improved. "The Home Office paperwork has got better, and the scientific community is more willing to stand up for what they do."

Tom Simonite and Jim Giles



HOW TO IMMUNIZE YOUR COMPUTER

Model shows viruses can be beaten at their own game.

www.nature.com/news

Animal-rights militancy exported to US and Europe

Some US and European animal-rights activists are adopting the illegal tactics of extreme UK groups, say police and groups monitoring protest activity. They note that just as incidents seem to be declining in Britain, they are rising elsewhere.

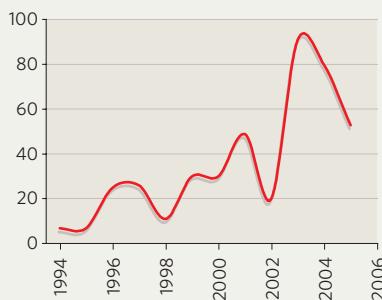
"When it comes to animal-rights extremism, there is a body of knowledge in the United Kingdom," says a spokesman for the UK police's National Extremism Tactical Coordination Unit, set up in 2004 to tackle illegal animal-rights activity. He says the UK scene has links with the United States and that "there has been an increase in extremist activity on mainland Europe directly related to that in Britain."

Mark Matfield is director of the European Biomedical Research Association, a London-based group that lobbies the European Union for better support and protection for researchers who use animals. He says that more sophisticated policing and tougher legislation have seen UK activity shift overseas.

"In the first half of 2005, there was an increase in illegal actions in Sweden, Switzerland, the Netherlands and Germany," Matfield says. "Much of it is either organized by British activists, or they have travelled abroad to get involved."

European police forces have detained British activists after illegal incidents, he says. Research laboratories are targeted indirectly, by attacks and threats to the property of their staff and those of

US ILLEGAL ANIMAL-RIGHTS PROTESTS



companies that work with them. And UK activist groups such as the Animal Liberation Front and Stop Huntingdon Animal Cruelty (SHAC) now have active branches in other European countries and the United States.

Carrie Wolinetz, spokeswoman for the Federation of American Societies for Experimental Biology, a Washington DC-based group that promotes biological research, draws the link explicitly. "The SHAC campaign and Animal Liberation Front started in the United Kingdom and were exported — unfortunately, because they were effective."

The fierceness of US animal-rights activism currently lags about five years behind the British, says George Goodno, spokesman for the Foundation for Biomedical Research, a Washington DC-based nonprofit organization that defends animal research and collects

information on illegal protests (see graph). "It definitely is ramping up," he says.

The United States became aware of the growing extremist activity in the country this September when Huntingdon Life Sciences, a research organization that conducts animal testing, was denied a listing on the New York Stock Exchange at the last minute. It was widely reported that this was owing to threats from SHAC, which proclaimed victory.

Parallel bills being considered in the US House and Senate would make it easier to prosecute animal-rights activists who cause economic damage to companies, academic laboratories and zoos; Britain passed similar legislation earlier this year. The US bills could mean a ten-year jail sentence for any activist who costs a company more than \$100,000.

John Lewis, a deputy assistant director at the Federal Bureau of Investigation with responsibility for counterterrorism, testified at a hearing on the Senate version of the bill. He said that activists kept mostly to a nuisance level of illegality — threatening phone calls, vandalism and raucous demonstrations in front of executives' houses.

Goodno has collected reports of 80 illegal actions by animal-rights activists in the United States in 2004, but notes that as the main tactic is constant low-grade harassment, these incidents are "the tip of the iceberg."

Emma Marris and Tom Simonite

SOURCE: FOUNDATION FOR BIOMEDICAL RESEARCH

TV show gives research lobbyist a rat's-eye view of laboratory life

To counter animal-rights activists, UK lobby groups that support animal research have launched their own media campaign.

In one reality television show, called *The Devil's Challenge*, the director of one such group is caged and subjected to procedures used in animal labs.

Simon Festing, director of the Research Defence Society in London, agreed to do the show, which was designed to test his belief in animal experiments. "We need to face the fact this is how the media works," he says, "and it's a good way to get our arguments out there."

Kept in a cage proportional in size to those used to house lab mice, Festing was subjected to a number of experiments. One recreated a test for pain, where rats are placed on a hotplate and the time until their feet twitch is recorded. Another, investigating wind chill, involved a dousing in water and a wind machine.

Other challenges brought him face to face with animal-rights activists and sent him to a primary school, where he tried unsuccessfully to persuade children to donate their cats for animal research.

Festing admits the experiences



Lab test: Simon Festing tries out life as a rat.

were difficult. "It did make me think harder about the welfare of lab animals," he says. But nevertheless, he remains a firm believer in the

need for animal research. *The Devil's Challenge* is broadcast on the UK digital channel More4 on 14 December.

T.S.

ON THE RECORD

“Unprotected sex with an infected individual is high risk regardless of whether the act is intended for procreation or recreation.”

Robert May, outgoing president of London's Royal Society, explains why the Vatican's ban on condom use encourages the spread of HIV.

“We are biologists and computer scientists, and what we do is just math. Math can't hurt you.”

Michele DeHart, head of the Fish Passage Center which, until recently, monitored salmon migration. Congress cut the centre's funding amid accusations that the institute was using data to promote an environmental agenda.

Sources: *The Age*, *Washington Post*

SCORECARD



Little penguins
A Tasmanian

conservation group has collected 15,000 tiny sweaters to help protect the world's smallest penguins from future oil spills.



Toucan beaks

A study at the University of California, San Diego, has found that the bird's beaks are remarkably well designed and could be used as models for stronger, safer car components.



Intelligent design class
The University of Kansas

has cancelled a course on intelligent design after the professor teaching it made disparaging comments about Christian conservatives, calling them “fundies”.

NUMBER CRUNCH

114 deaths per million people occurred in road crashes in 29 countries in the developed world during 2001.

0.293 deaths per million people were caused by terrorism each year in the same countries in 1994–2003.

390:1 is the ratio of road deaths to deaths from terrorism.

Source: N. Wilson and G. Thomson *Injury Prevention* **11**, 332–333 (2005).

LEE JAE-WON/REUTERS/CORBIS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Embattled: doubt has been cast on the validity of some of Woo Suk Hwang's work on human embryonic stem cells.

TV tests call into question cloner's stem-cell success

Woo Suk Hwang, the cloning researcher who last month admitted lying about the origins of human eggs used in his work, now faces questions about the validity of his scientific data.

In May, Hwang's team at the Seoul National University in South Korea reported it had established 11 embryonic stem-cell lines derived from the skin cells of patients (W. S. Hwang *et al. Science* **308**, 1777–1783; 2005). The experiment was hailed as a huge step towards the use of patient-specific cell lines in medicine. But Hwang has since sent *Science* two significant corrections to the published article.

And in a news programme on 1 December, the Seoul-based Munhwa Broadcasting Company (MBC) challenged the credibility of Hwang's data. Pursuing a tip-off, MBC procured from Hwang samples of five of the patient-specific cell lines and sent them, together with corresponding tissue samples, to an independent lab for DNA analysis. The programme reported that the DNA in one cell line did not match the tissue sample — as it should, if the lines were truly cloned from patient samples. DNA from four other cell lines could not be isolated.

According to Korean press reports, Hwang stands by the integrity of his science, but has yet to authorize independent tests, which could clear his scientific results.

There are many explanations for MBC's findings, such as improper matching of tissue and cell lines or contamination, says cloning expert Norio Nakatsuji of Kyoto University, Japan. But the DNA mismatch raises the possibility that existing or newly created embryonic stem-cell lines were substituted. “There would be no way to know from the paper whether the data were true or not,” he admits.

Donald Kennedy, *Science*'s editor-in-chief says that the journal is looking “very carefully into the history of this paper”, but warns against overreaction. “At the moment there is no reason to believe that any of this affects the scientific conclusions in the paper,” he says.

Last month, Hwang corrected a table in the original paper showing that all of the cells had passed a test to see whether they can divide into various cell types — a hallmark of embryonic stem cells. In fact, only three of the eleven lines had passed this test. And on 5 December, he notified *Science* that some images of stained cells, which supposedly represent different cell lines, were duplicates.

Alan Colman, chief executive of ES Cell International in Singapore and a member of the team that cloned Dolly the sheep, says it could all be down to “auditing error”. But some of the data are still very confusing, he adds.

Gerald Schatten, a co-author on the paper who is based at the University of Pittsburgh, has distanced himself from the article. *Science* has made an addition to the paper's supplementary information that describes the role of the University of Pittsburgh authors as limited to “the review and analysis of anonymized data and assistance in the preparation of this manuscript”. Although Schatten halted his collaboration with Hwang last month, he has stated that he believes the paper's conclusions are valid.

The confusion could easily be cleared up, says Colman, who organized an independent DNA analysis when sceptics raised doubts over the cells used to clone Dolly. “We were offended by allegations ranging from incompetence to fraud, but responded by clearing it up,” he says.

David Cyranoski



THE NATURE PODCAST
Listen to the stories behind
the hottest research in our
weekly show, available free
at [www.nature.com/
nature/podcast](http://www.nature.com/nature/podcast)

Avian flu centre put under threat of closure

The Indonesian government has ordered a US military research unit in Jakarta — a key player in the fight against avian flu — to cease all research by 31 December. If the unit closes, researchers say, it would be a major blow to efforts to control the avian flu outbreaks currently affecting humans and poultry across the country.

The US Naval Medical Research Unit No. 2 (NAMRU-2) was set up in Indonesia in 1970 under a bilateral agreement with the United States. But the Indonesian military has been opposed to the centre's presence since the imposition of a US arms embargo in 1999. This followed violence involving the army during Indonesia's withdrawal from the newly independent East Timor.

NAMRU-2 has been working with the country's authorities to improve their ability to monitor and diagnose avian flu (see C. G. Beckett *et al. Clin. Infect. Dis.* **39**, 443–449; 2004). Its crucial role is internationally recog-

nized, particularly as Indonesia, where human cases were first reported in July, accounts for the world's largest share of new cases.

Although the centre's original agreement with Indonesia expired on January 2000, until recently the government had turned a blind eye to its continued operation. But on 23 November, Indonesia's health ministry posted a memo on its website, dated 25 October, stating that "all NAMRU-2 activities must end by 31 December 2005", and not be restarted without a new formal agreement. The memo was addressed to all health agencies and hospitals.

Researchers associated with the centre, who requested anonymity, are taking this threat very seriously. "We take the memo as representing Indonesian government policy, and its statement is very clear," says one scientist.

The researchers had believed the future was clearer following a visit to Indonesia by the US health secretary Michael Leavitt in mid-October. Leavitt was on a whirlwind tour of

Asian countries to bolster belated US efforts to build an international coalition against avian flu. He visited NAMRU-2, and promised it US\$10 million in extra funding.

That the health ministry's memo should appear following Leavitt's visit points to internal government power-play, says one scientist, adding that on 16 November they had been privately assured by Dino Djalal, one of the Indonesian president's chief advisers, that the centre would be allowed to continue.

"If this is my government's policy then they must have their reasons," says Sardikin Giriputro, deputy director of the Sulianti Saroso Infectious Disease Hospital in Jakarta, which treats most of the country's avian flu victims. "I have enjoyed a lot of benefits and cooperation with NAMRU-2 — it would be a great pity if its activities were stopped."

The World Health Organization declined to comment on the development.

Declan Butler



US PUSHES LIMITS ON OZONE DESTROYER
Strawberry pesticide still essential, says farming industry.
www.nature.com/news

Climate talks edge towards twin-track future

MONTREAL

For many of the delegates from nearly 200 nations meeting in Montreal this week to discuss how to prevent climate change, the key question is how to get the United States to talk about limiting greenhouse-gas emissions.

As *Nature* went to press, the halls of the Palais des Congrès were rumbling with a report that the conference president, Canada's environment minister Stéphane Dion, would propose structuring climate-change strategy through the United Nations Framework Convention on Climate Change. Such an approach would be in addition to the Kyoto Protocol, which limits greenhouse-gas emissions for developed countries. The United States is a party to the convention, but not the protocol.

Such a strategy could free countries to pursue a range of options for fighting global warming, says Elliot Diringer, director of international strategies for the Pew Center on Global Climate Change, based in Arlington, Virginia. These options might include technologies such as carbon capture and storage, or fuel-economy standards for the automobile industry. "There's a rich array of thinking out here right now," says Diringer. "What we need is some window to introduce that thinking into the formal process."

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Slow puncture? Many believe the Kyoto Protocol got off to a wobbly start.

The gathering is the convention's first twin-track meeting. One track addresses the parties to the convention, and the other addresses the 157 nations that have ratified the Kyoto agreement. In the first week of negotiations, delegates finalized the rule book for the Kyoto Protocol, known as the Marrakesh Accords.

Australia and the United States are some of the holdouts from the Kyoto agreement. Developing countries such as China and India — some of the world's fastest-growing emitters — are not bound by emission caps, although they are parties to the protocol. The first commitment period for reducing emissions under Kyoto expires in 2012.

Some delegates welcome twin-track talks as a step towards a process outside Kyoto, which many feel has had an inauspicious start. The Clean Development Mechanism, for example,

an arm of the protocol intended to promote sustainable development in developing countries, has been slow to gain momentum. Few countries are expected to meet their current targets, making it less likely that they will take on bigger emissions cuts in the next commitment period.

But some countries could stall any forward-looking process. In a statement on 29 November, Harlan Watson, lead negotiator for the US delegation, said that

his country is opposed to discussing commitments beyond 2012. Watson reaffirmed the US position to engage instead in technological innovations on the home front and partnerships with smaller clusters of countries.

Such an approach does not please many. "They are clearly not moving forward on long-term cooperative action," says a spokesperson for the European Union presidential delegation, who asked to remain anonymous.

Many observers feel that a good finish for the conference, which runs until 9 December, would be a green light to discuss future options under the convention. "The most that we can hope for here is some decision to allow for parties to begin thinking about these options in a formal context," says Diringer. "It's an incremental step, but it's an essential step."

Amanda Haag

REUTERS/C. MUSCHI

Prospect of stricter timekeeping alarms US biologists

WASHINGTON

Proposed guidelines for federally funded US researchers, recommending more detailed accounting of how they spend their time, will, if approved, draw howls of protest from labs across the United States.

The office of the inspector general at the Department of Health and Human Services, parent department to the main US biomedical-research agencies, released the draft guidelines on 28 November.

The guidelines set out general principles for a programme to prevent and catch fraud, and identified three risk areas: time

reporting, sloppy accounting between grants, and not reporting financial support from other sources. Comments will be accepted until 28 December.

In the section on time reporting, the guidelines say, "Many researchers have multiple responsibilities — sometimes involving teaching, research, and clinical work — that must be accurately measured and monitored. In the course of a researcher's workday, the separation between these areas of activity can sometimes be hard to discern, which heightens the need to have effective timekeeping systems."

Most US researchers do not keep daily time logs, but estimate in advance what percentage of their time they will spend on a project. That estimate is used to determine how much of their salary and benefits the grant pays for.

The spectre of a punch-clock in the lab appals many scientists. "I think it would be a major impediment to American science," says Beth Levine, a specialist in infectious diseases at the University of Texas Southwestern Medical Center in Dallas.

"There is some very prescriptive language in there that gives us concern," adds Tony DeCrappeo, president of the Council on

Government Relations, a Washington DC-based association of research universities working to shape agency policy.

The guidance also calls for a compliance officer and compliance committee for each institution.

The guidelines are not mandatory, but the inspector general and the US Department of Justice might use them in judging fraud cases.

Glenn Baly, spokesman for the inspector general's office, says not to fret. "These are general guidelines. The details would be worked out by the agencies and organizations involved."

Emma Marris

German hostage was saving Iraq digs

Susanne Osthoff — the German archaeologist kidnapped in Iraq last month — was a lone force against looters, according to researchers in the field. They say that she was fighting to protect archaeological sites from the plundering that has been rife in the postwar chaos.

Field work in Iraq had to be abandoned after British and US troops invaded two years ago. And although the looting of the Iraqi National Museum in April 2003 received widespread media attention, the continuing loss of artefacts from more distant field sites is less appreciated. Osthoff was being “heroic” in trying to bring public attention to the crimes, says Michael Müller-Karpe, an archaeologist at the Roman–Germanic Central Museum in Mainz, Germany.

Hundreds of sites are being looted by ordinary citizens trying to make a quick profit, says Elizabeth Stone, an archaeologist at Stony Brook University in Long Island, New York. In a way, this widespread ransacking is worse than the looting of the museum, because the antiques being removed have never been catalogued, she says.

At first, little action was taken to protect the remote sites. But in the past few months about 1,000 Iraqi police have been deployed in the Dhi Qar province in the southeast of the country to look after them. “This has begun to have an effect,” Stone says. She has started to map the sites using satellite imagery, and hopes to make pictures available to Iraqi officials in Baghdad and in Nasiriyah, the province’s capital.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

In the field: archaeologist Susanne Osthoff (left) was trying to publicise the looting of digs in Iraq before she was kidnapped late last month.

Osthoff studied archaeology at the University of Munich, and in the late 1980s worked at a site in Isin, south of Baghdad. Now that field work is so dangerous, “a whole generation of students has not been involved in Mesopotamian archaeology,” laments Joan MacIver of the British School of Archaeology in Iraq, based in London.

Joanne Farchakh Bajjaly, a Beirut-based archaeologist, met Osthoff in April 2003 in Baghdad, when Osthoff was on her way to Isin

with a German television crew to report on the looting problem. Farchakh Bajjaly describes her as “a very courageous, strong woman”, saying such attributes are necessary for tackling a strong, well-organized mafia of antiques traders. Osthoff was aided by her fluent Arabic and her connections with a local tribe that helped to protect her, says Farchakh Bajjaly.

Osthoff’s kidnapping has raised the profile of the problem — but at a very high price, say archaeologists. As *Nature* went to press, she was still being kept hostage.

Andreas von Bubnoff

Cowrie study strikes a blow for traditional taxonomy

WASHINGTON DC

It may be too early to use a technology called DNA barcoding to speed the identification of species, says an analysis released last week.

DNA barcoding involves collecting and comparing genetic sequences from many species. Some proponents believe that bypassing the slower techniques of traditional taxonomy will identify unknown species and categorize the world’s biodiversity more

quickly. But a study by Christopher Meyer and Gustav Paulay of the University of Florida in Gainesville shows that barcoding works well only for species that are already much studied.

The scientists examined a database of marine snails called cowries, which have been studied since the nineteenth century because of their valuable shells. By analysing the sequence of a gene called *cytochrome oxidase 1* from a cowrie specimen, then comparing it

with the entire database of genetic sequences, the scientists correctly identified cowrie species with less than 4% error.

But Meyer and Paulay hit trouble when they reconfigured their database to examine how barcoding works for less well studied groups. They used a method pioneered by other barcoding scientists to pinpoint new species. This method compares diversity among members of the same species and diversity between different species.

They showed that in species that have been less studied, the barcoding method misidentified unknown specimens up to 20% of the time (C. P. Meyer and G. Paulay *PLoS Biol.* 3, e422; 2005).

Some scientists view the high success rate for known species as a triumph for barcoding. They also say that the field is very young, and is still working out the best ways to deal with organisms about which there is little information. “This was a very ambitious test for the technique,”

Chairman explains Europe's research council

M. BRITT HANSEN

The European Research Council (ERC), to be launched in 2007, will be the first Europe-wide granting agency for basic research. On 5 December, its Scientific Council announced the election of molecular biologist Fotis Kafatos as chairman. Alison Abbott takes this opportunity to quiz Kafatos, a lab chief at Imperial College London and formerly director of the European Molecular Biology Laboratory in Heidelberg, Germany.



First chairman:
Fotis Kafatos

create and supervise an efficient, flexible and responsive structure for our calls for proposals and the peer-review system.

Annual funding of €1.5 billion (US\$1.8 billion) has been requested for the ERC, but interest among researchers is so high that huge oversubscription is

predicted. How will you structure your calls so as to limit subscriptions?

The ERC will cover all fields of scholarship. One of our key tasks is to work out how to structure calls in a way that does not compromise this basic tenet. At the moment, no option is excluded. We will also think about a two-stage application procedure to relieve oversubscription.

How will the ERC peer-review system operate?

This is also something that we have to work out in detail. Certainly, it will take into account the international best practice.

What would happen if the Framework programme were to be squeezed hard during political negotiations, and the ERC given less money than requested?

It would not be worth establishing the ERC if it were poorly endowed — the funding has to be sufficient for it to work properly and make a difference. Personally, I think the absolute minimum would be €1 billion per year. ■

The ERC is designed to support the best basic research in Europe based on an independent peer-review system.

When will it start distributing money?

The political process for approving the seventh of the European Union's five-year Framework programmes for research will run through next year. The Framework's plan to create an ERC is secure, but it will have a legal basis only when this process is completed. We will be using this time for intensive preparatory work, so we should be ready to issue our first call for proposals very soon after that.

What is the role of the Scientific Council within the ERC structure?

The Scientific Council is the supreme body, and the agency will operate under our guidance. We will be assisted by a secretary-general who will have the experience to

B. D. COLE/CORBIS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

DNA barcoding identifies well known species such as this chestnut cowrie, but it comes a cropper on others.

barcode data," says Schindel.

Meyer stresses that the concept of barcoding remains solid. Indeed, his museum is a member of the barcoding consortium. But, he says, his study shows that barcoding will not supplant traditional taxonomy. Meyer says he hopes his study will help convince funding agencies to support old-fashioned taxonomy — a discipline that seems to be slowly disappearing — as well as DNA barcoding. Erika Check

says David Schindel, executive secretary of the Consortium for the Barcode of Life, a group of 93 barcoding practitioners hosted by the Smithsonian Institution in Washington DC. "Chris and

Gustav have done a very good job of presenting the possible procedural problems that barcoding will encounter, and we are working out protocols for more sophisticated analysis of

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Making progress: Robert Klein (left) is optimistic about the future of California's stem-cell initiative.

Court ruling offers hope for California stem-cell project

The institute at the heart of California's \$3-billion stem-cell research programme is claiming an early victory in its legal struggle to distribute its funds.

Set up in the wake of a state referendum last November that backed the project, the California Institute for Regenerative Medicine (CIRM) in San Francisco has since been beset by lawsuits that have stopped it handing out cash to researchers.

But in a ruling handed down on 29 November in the Superior Court of

California in Alameda County, Judge Bonnie Sabraw dismissed many of the arguments made by the initiative's opponents, who say it is unconstitutional and should be voided.

The ruling "provides the CIRM with a strong basis for moving forward successfully in this case", says Robert Klein, chair of the institute's Independent Citizens' Oversight Committee. As *Nature* went to press, Sabraw had scheduled a follow-up hearing for 6 December.

Meanwhile, the UK Stem Cell Initiative is seeking at least £350 million (US\$610 million) in additional government funding for stem-cell research over the coming decade. The government has pledged an extra £50 million over the next two years.

Sweden becomes home to neuroscience data centre

The Karolinska Institute in Stockholm, Sweden, is to host a new international facility aimed at helping neuroscientists share and analyse data.

The International Neuroinformatics Coordinating Facility (INCF) will assemble a network of neuroscience databases. Plans for the facility were approved last year at a ministerial science meeting of the

Organisation for Economic Co-operation and Development, and its Global Science Forum has since reviewed bids by several countries to host the headquarters. The decision was finally revealed at the inaugural meeting of the INCF's country representatives on 28 November in Paris.

NIH proposes grant cuts for young researchers

The US National Institutes of Health (NIH) is considering cutting back its support for young researchers. At a 30 November town-hall meeting in Maryland, agency officials heard responses to a series of proposed funding cuts for graduate students and postdocs on research fellowships.

In the current budget environment, officials said, the agency would be unable to sustain the 17,000 grants it gives out each year to young researchers. In compensation, it proposed a cap on tuition reimbursement; a fixed tuition allowance of \$16,000–18,000; or a cut in the total number of grants.

University officials said they were unhappy with all three proposals, as they would force universities to pay towards supporting young researchers. The NIH expects to make a decision by next spring.

Cash boost gives genetics institute a broader remit

The billionaire philanthropists whose \$100-million donation launched the Broad Institute in Cambridge, Massachusetts, have doubled the value of their gift.

On 30 November, Eli Broad and his wife Edythe announced that they are giving the 18-month-old genomics research centre an additional \$100 million.

Like the initial funding, the new cash must be spent over the next ten years, committing the institute to spending \$20 million each year. This is just a fraction of the centre's annual research budget of about \$100 million, most of which comes from government grants. The new money will be spent on wide-ranging interdisciplinary projects that typically go beyond the remit of federal funding.

Prize applauds research from climate to pulsars

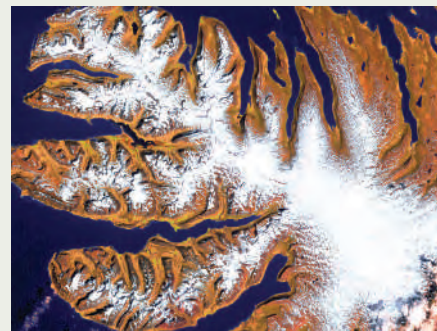
Five international research teams share this year's €1-million (US\$1.2-million) Descartes Prize for Research from the European Commission, awarded for

Power problems put satellite out of the picture

Landsat 5 is in trouble. One of the two remaining craft in the US programme for taking images of Earth's surface (such as the picture of fjords in Iceland, right), the satellite suffered a hardware glitch on 26 November that led to it being temporarily shut down.

The problem lies in the solar array that powers the craft. In January, the drive mechanism that ensured Landsat 5's main array tracks the Sun stopped working; now its back-up drive is showing similar signs of failure. If this does mean the end for the 21-year-old satellite, it will still have done extremely well — it was built to last just three years.

The other satellite active in the programme, Landsat 7, was launched in 1999 but suffers from a malfunction that means it must take at least two images of a scene to ensure that the whole picture is sharp (see *Nature* **423**, 907;



NASA

2003). No other satellites are planned for the 32-year-old Landsat programme, although Congress has given the agencies responsible for running the project a deadline of next March to come up with a new way to keep the data flowing.

excellence in collaborative research.

The EXEL team was recognized for work on metamaterials, which can bend electromagnetic radiation in unusual ways and are potentially useful for focusing radio waves or making 'perfect' lenses.

PULSE was rewarded for its studies of pulsars, including the discovery of the first double pulsar system. The EURO-PID project won for its work on more

than 130 rare diseases called primary immunodeficiencies, which leave sufferers open to infection and autoimmune disorders.

The European Social Survey was selected for its attempts to track social change in European Union member states, and the CECA group received plaudits for its programme to monitor sea-ice cover and other climate-related changes.

RUNNING THE RED LIGHT

A trial drug encourages cells to ignore the signs that stop them making faulty proteins. Sound dangerous? **Claire Ainsworth** discovers that it could be a cure for genetic disease.



On 28 November, two patients in an Israeli hospital downed a vanilla-flavoured shake in the hope of curing their cystic fibrosis. The drug in the shake is meant to get their cells to ignore the genetic mutation that makes their lives a misery. Remarkably, if this trial works, the very same drug might be used to help patients with a huge range of genetic diseases.

At the moment treatments for cystic fibrosis concentrate on the condition's symptoms, such as lung-clogging mucus, rather than its underlying cause — a faulty gene. Advocates of gene therapy aim to correct this by delivering working copies of the cystic fibrosis gene to a patient's cells, but results have been mixed so far.

The drug in the shakes, PTC124, takes a different approach, concentrating neither on the gene itself nor on the symptoms it causes, but on the process that links the two. It is supposed to work on the system that translates genetic information into proteins, coaxing it into ignoring a specific sort of genetic defect. The drug cannot offer aid to all those with cystic fibrosis, because it is expected to help with only a particular subset of mutations. But if it works for a few people, it should also work against some of the other diseases in which single genes are damaged. It is estimated that one-third of people with an inherited genetic disease have the sort of mutations that PTC124 and similar treatments might get cells to ignore.

PTC124 is aimed at thwarting a cellular process called nonsense-mediated decay, or NMD. Research into this phenomenon has taken off only in the past ten years, but it is already offering potential therapies and insight

into the mysteries of information processing in the cell¹. Some of the scientists involved hope that NMD will shed light on classic puzzles, such as why creatures with similar genetic make-up can have different physical characteristics and why our genome has evolved the way it has.

In the first step towards producing a protein from a gene, the cell makes a complementary RNA copy of the gene's DNA sequence. This copy, called messenger RNA, is packaged and processed and sent off to the cell's ribosomes — molecular machines that read RNA and put together proteins according to the instructions contained in the sequence. Each sequential group of three genetic letters, or codons, in the RNA tells the ribosome to add a particular amino acid to the protein that it is constructing. If the gene is damaged in some way, the RNA will be too. If a piece of DNA is missing, some codons will be lost; if it is garbled, the codons will be as well, and the resulting amino-acid chain may not be a properly functioning protein.

But not all codons signify an amino acid. Some, the 'stop' codons, mark the end of the gene's protein-coding sequence; they tell the ribosome that the chain of amino acids it has been making should come to a close. Some mutations can cause a stop codon to appear in the middle of a messenger RNA, and it is these mutations that cause the NMD response.

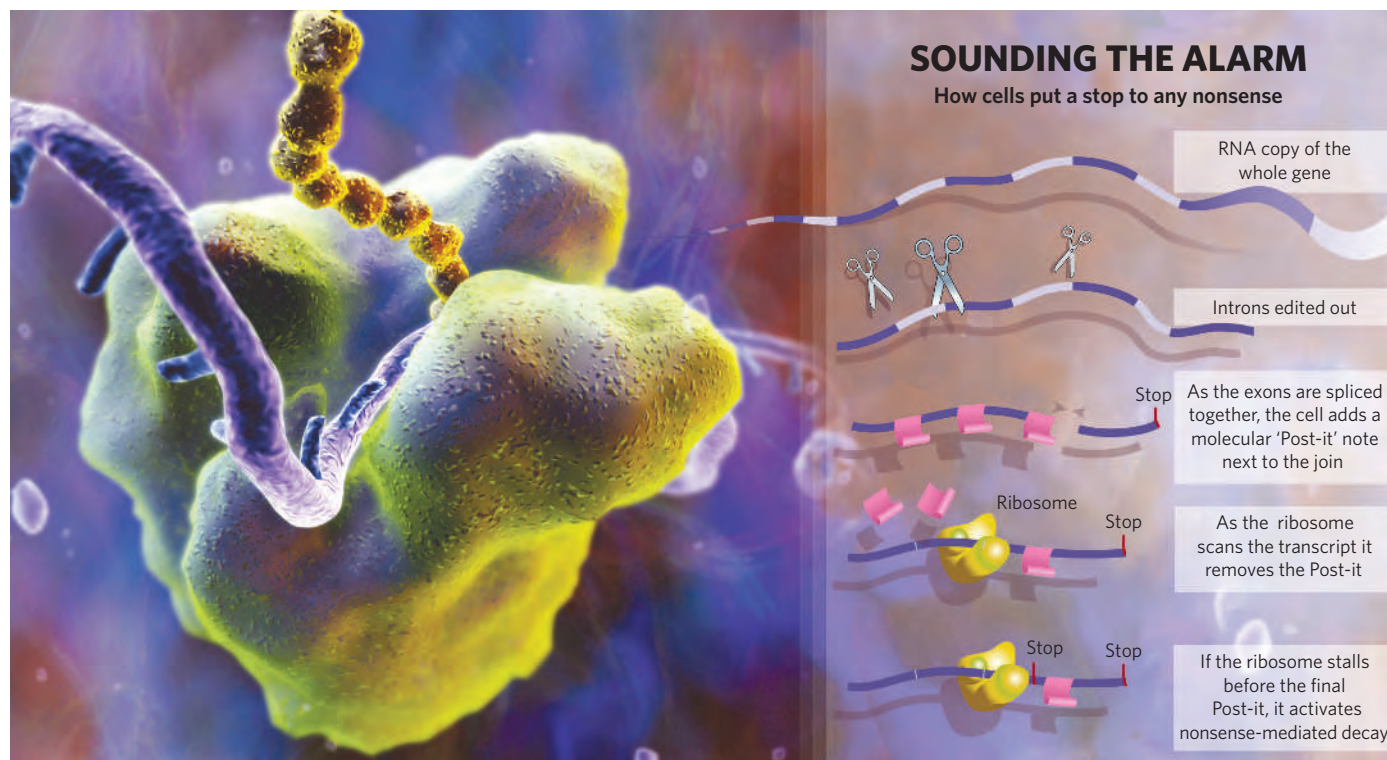
In the late 1970s, scientists noticed something odd about a subset of patients with a potentially fatal inherited anaemia called β -thalassaemia. This disease affects β -globin, one of the subunits of haemoglobin, which carries oxygen in the blood. In some forms of the disease, patients make faulty β -globin; in

others they don't seem to make the protein at all. Lynne Maquat, then a postdoc at the University of Wisconsin, Madison, was studying the bone marrow of patients that apparently made no β -globin and found that the messenger RNA for the protein disappeared unusually rapidly from their cells². Maquat and others sequenced these patients' genes and found stop mutations in them. It seemed as if the extra stop codons were triggering a quality-control process that disposed of the faulty messages. But no one could work out how a cell tells the difference between a normal stop codon and one caused by mutation.

Don't shoot the messenger

A key clue came from a set of patients with a rare form of β -thalassaemia. These patients made a shortened version of the protein — the version you would expect to see if the cell was allowed to read the faulty RNA and simply finished the protein when faced by a premature stop codon. The thing these rare cases had in common was that their extra stop codons all lay towards the end of the β -globin gene. Subsequent work by Maquat, now heading a team at the University of Rochester, New York, and by other groups has shown that the explanation lies in the way that messenger RNA is patched together from its gene³.

Not all the DNA in a human gene codes for amino acids. The bits of the gene that carry instructions for making the protein are called exons; the intervening bits are called introns. The introns are cut out of the RNA copy of the gene, and the remaining exons are spliced together. These splicing points have marker proteins stuck near them, like Post-it notes (see graphic).



Researchers think these notes act as signposts, telling the ribosome where it is as it moves along the RNA, explains Elisa Izauralde, who works on NMD at the European Molecular Biology Laboratory in Heidelberg, Germany. If the first ribosome that encounters the messenger RNA makes it past the final Post-it, the cell will be satisfied that the RNA is up to scratch, and allow other ribosomes to work on it. But if the scouting ribosome finds a stop codon before that point, alarm bells ring and the faulty message gets shredded by NMD. In the rare thalassaemia cases, the RNA survived to produce protein only because the stop codon appeared too late in the sequence to trigger the decay response.

The protein fragments that premature stop codons would produce if left to their own devices could do various types of damage; for example, they might stick to other proteins and stop them working properly. NMD gets rid of this problem. In some cases this limits the harm done by mutations with which we are born; in others it may lessen the effects of mutations we later acquire. The development of cancer often involves the appearance of stop codons in genes that would otherwise protect us from the disease, and it may be that the efficiency of a person's NMD response affects their predisposition to cancer. "It's an attractive speculation, and may well turn out to be true, but it is not yet proved," says Philip Anderson, a biologist who works on NMD in the *Caenorhabditis elegans* worm at the University of Wisconsin.

Still, NMD is a double-edged sword. In β -thalassaemia, shortened versions of the β -globin protein really would do damage, and it makes sense to silence the gene through NMD.

But shortened versions of the cystic fibrosis protein do not cause disease. "It's the other side of the coin," says Andreas Kulozik, a paediatrician at the University of Heidelberg who studies the decay response and its role in blood diseases. "If these truncated proteins were actually made, they would do some good." The same is probably true in many other conditions; not all protein fragments are harmful, and some help.

Deliberate mistakes

Researchers have not yet worked out how to control NMD. But what they can do is smuggle nonsense mutations under NMD's radar, and fool the cell into producing a protein that the decay response would prohibit. In a neat twist, they learned this trick from an antibiotic commonly used to treat the lung infections that plague cystic fibrosis patients.

Gentamicin belongs to a family of antibiotics called aminoglycosides that gum up the ribosomes of bacteria, making them prone to misreading messenger RNA. These befuddled ribosomes will often mistake a stop codon for one specifying a run-of-the-mill amino acid and 'read through' it, adding that amino acid in its place. Researchers wondered whether they could pull off the same trick in human cells and so tiptoe past the NMD machinery. If the ribosomes were to read through instead of stalling, the decay process wouldn't come into play, and

nearly normal proteins should be produced.

In 1997 a team at the University of Alabama, Birmingham, announced that it had managed to get human cells in culture to read through a faulty cystic fibrosis gene with the aid of gentamicin⁴. Two years later, a team led by Lee Sweeney of the University of Pennsylvania, Philadelphia, showed that the strategy could work in living animals⁵. Studying mice with nonsense mutations in the dystrophin gene, which is mutated in Duchenne muscular dystrophy, Sweeney's team found that those dosed with gentamicin produced full-length dystrophin protein, although much less of it than normal mice do. Movement did less damage to the muscles of treated mice than untreated mice.

The papers caught the eye of Eitan Kerem, a cystic fibrosis specialist at the Hadassah University Hospital in Jerusalem. He and his team had noticed that some of their patients with cystic fibrosis improved to an inexplicable extent when given gentamicin to inhale for lung infections, relapsing badly when the drug was withdrawn. "In some patients, we used to call it a magic drug, because they improved remarkably," he recalls. "We always said: there must be something else in this drug." And they were in a particularly good position to test the idea that gentamicin was suppressing NMD. In most populations, nonsense mutations are the cause of cystic fibrosis in just 2–5% of cases. But owing to an ancient genetic accident and centuries of intermarriage, some 60% of Ashkenazi Jews with cystic fibrosis carry a nonsense mutation.

Kerem's team gave gentamicin nose drops to 19 cystic fibrosis patients who had stop-codon mutations. The drug restored protein function

"We used to call gentamicin a magic drug, because some patients improved remarkably."

— Eitan Kerem

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

in the nasal membranes of 90% of them⁶. In itself this did not lead to a clinical improvement, but Kerem was encouraged enough to start planning to administer the drug by inhalation at higher doses than those used to treat infections.

At that point he was contacted by Stuart Peltz, founder and chief executive of a biotechnology company called PTC Therapeutics, based in South Plainfield, New Jersey. By looking for small molecules that interfered with ribosomes, Peltz's company had developed a drug that, like gentamicin, encouraged read through of nonsense mutations. At the 2004 North American Cystic Fibrosis conference in Anaheim, California, Peltz revealed unpublished data showing how PTC124 partly restored protein production and normal physiology in a mouse model of cystic fibrosis. The drug has now completed the first phase of safety trials in healthy volunteers, and unlike gentamicin, which can cause kidney failure and deafness, its immediate side effects seem rare and mild. It is also readily absorbed through the gut, making it easy for patients to take, says Peltz.

Phase II efficacy trials on adult patients have begun in both the United States and Israel. Cystic fibrosis patients are already trying the drug — with 24 involved in Kerem's trial — and patients with Duchenne muscular dystrophy are being recruited. The results, due in mid-2006, should show whether PTC124 offers clinical benefits. If it does, the company hopes to try it on other conditions, possibly including cancers in which nonsense mutations play a

role. The researchers are aware, however, that letting cells make proteins that NMD would normally prohibit may have long-term risks, and these need to be weighed against the advantages of drug treatment. "I suspect that you don't want to tamper with that too much," says Kulozik. "But for certain conditions, tampering a little bit is likely to be beneficial."

The detailed workings of NMD are still unclear. "We think we know most of the players in the game, but we don't know how they play with each other," says Matthias Hentze, a biologist at the European Molecular Biology Laboratory.

Switched on

In the future, researchers hope to control the decay response in a more sophisticated fashion, says Maquat. And it is a reasonable goal. There are signs that the some of the proteins involved in the decay can receive signals from inside or outside the cell, which raises hopes that the system could be turned on and off. Kulozik's team, together with Hentze's, recently discovered at least two distinct switches for the NMD response in mammalian cells⁷.

Evidence such as this suggests that NMD is much more than a mere cellular janitor. Researchers know, for example, that NMD proteins are needed for efficient translation of normal messenger RNAs⁸. It could also be involved in controlling gene expression. The ability to splice exons together in different ways lets cells make more than one protein from the same gene. This is a useful way to get the most out of a genome, but

Clear the lungs: most treatments for cystic fibrosis simply tackle its symptoms.

it is also a rather messy one — likely to produce some RNAs with stop codons in them. A nonsense-mediated response reduces the risk that these stop codons will cause harm. And the interplay of NMD and the stop codons introduced by splicing may, in fact, be a way of controlling how and when genes get turned into proteins⁹. "As time goes on, there are more and more examples of 'purposeful' targets of NMD," says Anderson, citing his work on *C. elegans*. "These are not errors of gene expression, they are normal aspects of it."

Michael Lynch, a biologist at Indiana University in Bloomington, suggests that NMD may have helped to spread introns through genomes. Using a mathematical model, Lynch discovered that introns are far more evenly spaced in vertebrate genomes than would be expected by chance. Introns in organisms such as yeast and fruitflies, which do not seem to use the Post-it system to trigger NMD, are rarer or more randomly distributed. Lynch speculates that introns and the decay response scratched each other's backs through evolution: NMD meant that errors caused by multiplying introns were minimized, while introns that spaced themselves evenly in genes acted as regular signposts for ribosomes, making NMD work better. "But who was driving whom is really an open area of research," he says.

Anderson agrees that the decay response is a force to be reckoned with in evolution. By masking the effects of mutations that would otherwise reduce a creature's ability to survive, NMD gives genomes much greater freedom to experiment and evolve.

What is more, Anderson's team has shown that differing versions of the same gene behave differently, depending on the presence or absence of NMD machinery¹⁰. It is this idea that excites medics such as Hentze and Kulozik, who wonder whether differences in NMD could influence how badly people are affected by inherited diseases.

As well as saving lives, it seems that understanding NMD could help us to explain some of biology's biggest questions. But researchers still need hard data from molecular biology to make these theories stick. "It's easy to make these speculations, and very entertaining," says Anderson. "It's damned hard to prove, but to me it makes a lot of biological sense." ■

Claire Ainsworth is a senior News and Features editor for Nature.

- Holbrook, J. A., Neu-Yilik, G. N., Hentze, M. W. & Kulozik, A. E. *Nature Genet.* **36**, 801–808 (2004).
- Maquat, L. E., Kinniburgh, A. J., Rachmilewitz, E. A. & Ross, J. *Cell* **27**, 543–553 (1981).
- Maquat, L. E. *J. Cell Sci.* **118**, 1773–1776 (2005).
- Bedwell, D. M. et al. *Nature Med.* **3**, 1280–1284 (1997).
- Barton-Davis, E. R., Cordier, L., Shoturma, D. I., Leland, S. E. & Sweeney, H. L. *J. Clin. Invest.* **104**, 375–381 (1999).
- Wilschanski, M. et al. *N. Engl. J. Med.* **349**, 1433–1441 (2003).
- Gehring, N. H. et al. *Mol. Cell* **20**, 65–75 (2005).
- Wilkinson, M. F. *Trends Genet.* **21**, 143–148 (2005).
- Mendell, J. T. et al. *Nature Genet.* **36**, 1073–1078 (2004).
- Calì, B. M. & Anderson, P. *Mol. Gen. Genet.* **260**, 176–184 (1998).



Back on track?

Next June, a \$1.4-billion neutron-scattering facility will come online in the United States. **Karen Fox** finds out whether this machine really can breathe fresh life into the ageing Tennessee lab that is its home.

Oak Ridge National Laboratory is more than 60 years old and, until recently, it looked that way. Despite its track record in nuclear research, the host of wildlife that wanders on campus, and the pretty sunsets over the rolling hills of eastern Tennessee, it looked unlikely to entice the energetic people who are the lifeblood of any great laboratory.

Salvation may be at hand. The Spallation Neutron Source (SNS), the largest scientific facility to be built in the United States for a decade, will become operational at Oak Ridge by next June. The \$1.4-billion machine will generate neutron beams by firing high-energy protons at liquid mercury. Hundreds of visiting researchers are expected to descend on the laboratory and use these beams to probe the structures of molecules and crystals.

"What this brings to the table is the opportunity to do an entirely new class of experi-

ment," says Jack Rush, who retired earlier this year as director of neutron scattering at the National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland.

Researchers around the world will be keen to assess the popularity and capabilities of the new facility. Once there would have been no question about the usefulness of such a resource. But in the two decades since the US Department of Energy (DOE) first planned a neutron facility at Oak Ridge, the options available for mapping molecular and crystalline structures have widened. Facilities that use X-rays to probe these structures, such as the Advanced Photon Source at Argonne National Laboratory in Illinois, have become vastly more powerful. Meanwhile, the number of researchers who work with neutrons has declined as several ageing neutron-source facilities have closed.

Despite this, Japan is building a research

facility similar to the SNS at Tokai, which will be ready in 2007. European plans for an even more advanced spallation source hit political setbacks a couple of years ago; their proponents are keen to revive them as a way of preserving the continent's long-standing lead in neutron science.

Building excitement

At Oak Ridge, the new neutron source is generating a palpable buzz. The facility managers are aware that many potential users of the technique have learned to live without it. So to bring them to a site far from the universities where they work will require exceptional management and technical support.

"You can have a philosophy of 'if you build it, they will come,'" says Paul Butler, team leader for NIST's small-angle neutron-scattering instrument. "But in my experience that doesn't work. You have to do more." Butler is

on the SNS users group committee, formed when construction began in 1999, to make sure the neutron source goes that extra mile to meet scientists' needs.

Oak Ridge managers know they face a challenge. They anticipate that just a few hundred researchers will use the facility during its first two years of operation, while its neutron output slowly ramps up to its full potential. But they expect this to build up to 2,000 users a year by 2015, as researchers currently accustomed to using advanced X-ray sources begin to be converted to the subtler and sometimes complementary charms of neutrons.

Supporters of the neutron source say that it does things that other mapping tools can't do. X-rays bounce off electrons, and so scatter much more spectacularly from heavier elements that have many electrons than from lighter ones, such as hydrogen, which has only one. It has been estimated that, as a result, the placement of about half of the hydrogen atoms in published protein structures derived from X-ray studies are not known. Neutrons interact directly with nuclei, making lighter atoms easy to identify.

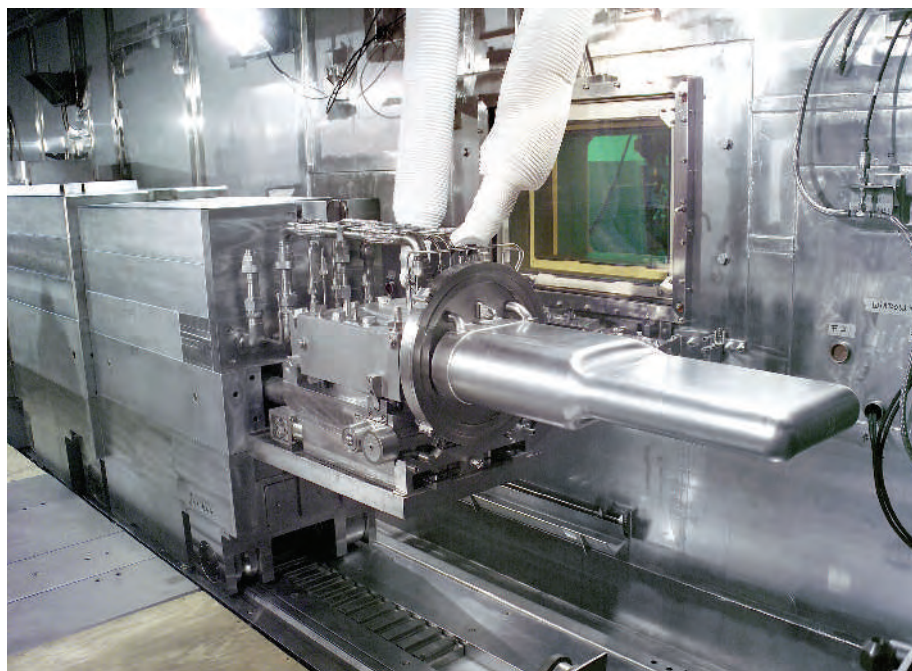
Structured approach

Dean Myles, head of structural biology at Oak Ridge, points out that another advantage of dealing with nuclei, rather than electrons, is the ability to distinguish between different isotopes of the same element. This means that researchers can, for example, use deuterium — a heavy isotope of hydrogen — as a marker for the position of a particular atom in a molecule or structure. "I liken it to a black cat in a snow field," says Myles. Because neutron sources can map structures over time, a molecule labelled with deuterium could, for instance, be watched as it wanders across the surface of a sample that mimics a cell membrane. Oak Ridge is keen to teach its users the chemical tricks necessary for such work, and the lab is also growing bacteria in deuterium-rich media so that they produce deuterated proteins.

Neutrons can also probe magnetic moments in solids and phenomena such as high-temperature superconductivity. Theories explaining superconductivity can be tested by mapping the position and movement of oscillations of the magnetic moment at a range of different atomic energy levels, Rush says. And the SNS should be able to collect these data some ten times more quickly than existing neutron facilities, owing to its high neutron flux.

This high flux level — up to 10^{17} neutrons per square centimetre per second — will also allow users of the facility to extract useful information from smaller samples. This, says Rush, is a valuable capacity for people studying things in short supply, such as proteins or newly developed polymers.

The facility also incorporates specialized equipment to cater for different research needs. One instrument will place samples under extremely high pressure, helping



OAK RIDGE NAT'L LAB.

On target: mercury held inside this vessel will be hit by high-energy protons to generate a beam of neutrons.

planetary scientists who want to model the hydrogen-rich interior of Jupiter. Until now, it has been a struggle to reach pressures above 25 kilobars simply because of a lack of beam intensity at neutron sources, says Richard Nelmies, who specializes in high-pressure neutron science at ISIS, the British neutron facility near Oxford, which is currently the most powerful spallation source in the world.

For Oak Ridge — the largest of the DOE's civilian laboratories — the new facility provides a badly needed opportunity to regain scientific momentum. "It's a little bit of an engine you get rolling," says Jeff Wadsworth, Oak Ridge's director. "It generates an optimism that feeds on itself."

Face lift

Oak Ridge was built in 1943 to produce uranium and plutonium for the Manhattan Project, and a major new facility hasn't been added since the high-flux isotope reactor was built in 1966. "There was first-class research and great people, but it looked like a decrepit 1950s lab," says Thomas Mason, who joined the SNS project at Oak Ridge in 1998, becoming the lab's

source reactor — allocated to Oak Ridge. But in 1995 Congress halted the \$2.9-billion project just before construction began.

The DOE decided to build a less expensive, accelerator-based neutron source instead: the SNS. Developed as a joint project between a number of the department's laboratories, including Los Alamos in New Mexico and Brookhaven in New York state, it was decided to locate the facility at Oak Ridge, in the home state of then-vice-president Al Gore.

After construction began, Oak Ridge's management contract was taken over by the University of Tennessee and Battelle, a contract research organization based in Ohio. The new management team has been working hard to secure extra investment for the lab: for instance, it has borrowed \$115 million from private banks to build associated infrastructure, including a new centre for computational sciences. On the back of that, Oak Ridge has won leadership of a large DOE supercomputing initiative.

"There is a substantial amount of risk that goes along with the debt," says Wadsworth. "But we believed it would help to attract more contracts and to grow our business, and so far we've been successful."

Ultimately, Oak Ridge will measure its success by its ability to attract world-quality researchers — both as visitors and as staff. "You put enough bright people together and interesting things happen," says Butler. "There are all these buildings going up on the hill, and then you put in these people and resources, and add 1,500 users with all their new ideas." In these gentle southern hills, he predicts, "it's going to be a melting pot of ideas, bubbling away".

Karen Fox is a science writer based in Washington DC.

"You put enough bright people together and interesting things happen." — Paul Butler

associate director three years later. "People weren't working with state-of-the-art facilities."

Next year's opening is the culmination of a prolonged struggle to rectify that. Back in 1984, a National Academy of Sciences panel recommended the distribution of various scientific facilities to different DOE labs, with the largest one — a proposed advanced neutron-

Screen test

A new technique could allow doctors to spot hundreds of potential genetic problems in unborn babies. But is it too soon to put it to use? **Erika Check** finds out.

Three years ago, a doctor told Debbie Sukin that her son had a rare and serious genetic disease called Angelman's syndrome. The diagnosis meant that her son, then just one year old, would face tremendous physical and mental challenges for the rest of his life. After the diagnosis, Sukin went to see Arthur Beaudet, a leading expert on the syndrome. Beaudet tested Sukin and her husband, and found that neither of them was carrying the genetic fault that had caused her son's disease — the condition had arisen spontaneously in the unborn child. Because there was no reason to suspect a problem, Sukin did not have any genetic tests performed during her pregnancy.

Beaudet, a geneticist at the Baylor College of Medicine in Houston, Texas, felt that Sukin's story represented a broader problem in medicine. He could see a widening gap between geneticists' growing understanding of the roots of disease and their inability to detect those diseases in the womb. The main problem is that prenatal tests can only catch genetic problems if doctors know to look for them. For Beaudet, this was simply not good enough.

To tackle the problem, Beaudet's lab has developed a way to test for more than 150 chromosomal abnormalities using a single package that costs just under US\$2,000 a go. Over the past year, Beaudet has tested the technique in an unpublished clinical trial in 98 women who were at high risk of having babies with genetic problems. He was so convinced by the results that this August his clinic became the first in the world to offer the prenatal test. Beaudet predicts that the test will change the face of medicine. "This is going to cause a world revolution in prenatal and perinatal care," he claims. Beaudet's claims are perhaps optimistic — for now, the test will only be used by women who can already afford genetic screening.

Other doctors and scientists agree that the technique has huge potential, but they worry that it is too soon to use it in unborn children. They say it could pick up genetic features that are difficult to interpret, causing extra anxiety

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Hard to spot: current tests for genetic defects in fetuses rely on doctors anticipating what they might find.

for parents-to-be. In some cases, the results mean that additional tests must be done to examine DNA from both parents, which could reveal instances of what geneticists delicately call 'non-paternity'. And others fear that the test will give parents more opportunities to terminate fetuses with undesirable traits.

Beaudet's technique is the latest development in a long line of tests that examine chromosomes in the fetus. Until recently, doctors could detect only large deletions, copies or rearrangements of chromosomes, such as an extra copy of chromosome 21, which causes Down's syndrome. Now, modern techniques such as fluorescence *in situ* hybridization, or FISH, can pick up much smaller changes.

A wider net

Beaudet's test has two powerful advantages. It picks up even smaller changes than the FISH test, and unlike FISH it simultaneously screens hundreds of chromosome areas that have been linked to disease. Doctors don't usually check for mutations in all of these areas because most of the mutations are extremely rare accidents that occur during development. Any abnormalities that show up can then be investigated further. "Our system allows you to do every known

FISH test in the world at once," says Beaudet.

The new test uses microarray-based comparative genomic hybridization, or array CGH. This is based on the principle that every cell should have two full copies of its DNA — one from the mother and one from the father. The test scans for regions of fetal DNA that deviate from this pattern because they contain too much or too little DNA. These aberrant patterns correspond to regions of the genome that are either copied or deleted, and could therefore cause disease.

Although array CGH is still very new, a few labs in Europe and the United States are already using it in the clinic. But in most instances, they use it only to look for genetic problems in children or adults who have unexplained mental retardation¹⁻³. And, unlike Baylor, these labs believe that it is too early to use the test as a prenatal screening tool.

Their hesitation stems from a new understanding of human genetic diversity⁴. Researchers are finding that individuals who seem perfectly healthy often carry deletions and duplications of certain genes. The Baylor test looks for mutations across long stretches of DNA, and it could be difficult to predict the consequences of deletions and duplications of these regions in a fetus. "Until we have a much better understanding of what normal variation is, it is dangerous to launch into clinical testing in the prenatal context," says Martin Bobrow, a medical geneticist who is working with a group at the Wellcome Trust Sanger Institute near

"The big question is, where are we going to go with all of this?"
— Dorothy Mitchell-Leef



AGAPITOSANCHEZ@BCM2005

Informed choice: Arthur Beaudet discusses his prenatal genetic screening test with an expectant mother.

Cambridge, UK, on developing pre- and post-natal diagnostic tests based on array CGH.

Bobrow and others are also concerned about the test's power to uncover ever more detailed information about a developing baby. For instance, an old version of one genetic testing device scanned regions of the Y chromosome that allow a man to make sperm. If found in a fetus, defects in this part of the chromosome might signal that a baby boy is destined to grow into an infertile adult, perhaps leading some parents to think twice about keeping the baby. But, as Bobrow says: "Very few doctors would want to be involved in terminating a pregnancy on the basis of male infertility."

"The big question is, where are we going to go with all of this?" asks Dorothy Mitchell-Leef, a fertility doctor at Reproductive Biology Associates in Atlanta, Georgia. "I doubt if there's anyone alive today who is a perfect example of a healthy individual and has absolutely no disease." That makes it difficult to know what genetic changes really mark a fetus as abnormal, she notes.

Baby steps

Beaudet is aware of these pitfalls. But he argues that genetic counsellors have been dealing with these issues since the 1970s, when prenatal testing began. "We have always faced findings of uncertain significance in prenatal diagnosis," he says.

He adds that the results of Baylor's year-long preliminary study, described at the American Society of Human Genetics meeting on 26 October in Salt Lake City, Utah, have been reassuring. The test turned up no abnormal findings in the vast majority of pregnancies. In five of the 98 women taking part in the trial,

the test uncovered obvious abnormalities — four were the tell-tale signature of Down's syndrome. In nine other cases, the test detected a genetic hiccup that wasn't known to be associated with any disease, but the variation was also found in one of the baby's healthy parents, so doctors assumed it was not dangerous.

Only once did the test turn up a variant that was not associated with disease and was also not found in either of the baby's parents. In future cases, Beaudet says, the group would like to perform a paternity test in these situations, to be sure it is looking at the right parental DNA. He admits that this raises issues of its own. "It's relatively new territory for prenatal diagnostics to have to look at data from both parents to interpret data on the fetus," he says.

But, he adds, the Baylor group will refine its methods as it learns which DNA variations are harmless to the fetus. And both Beaudet's team and other groups have performed studies to prove that the technology actually detects genetic changes that lead to disease^{5,6}.

Beaudet's prenatal work has followed strict ethical guidelines. But there is a legitimate

worry that less scrupulous operators could develop the test to screen for genetic variations associated with desirable traits, such as height or IQ. Although more rigorous than in most other parts of the world, US regulation of genetic tests is still somewhat patchy. So some doctors are calling for self-regulation. Mitchell-Leef believes that medical societies should set policies now on what sorts of conditions should be tested in embryos and fetuses.

Beaudet and his group agree that such issues are important. But just as important, they say, are the voices of the pregnant women who have requested the test so far. This January, for instance, Anca Segall, a biologist at San Diego State University, had her unborn child tested. At 44, she was anxious to know upfront if her child had any major problems. "You really have to think about whether this knowledge is important for you, and it was for us," Segall says. Her test raised no red flags, and she gave birth this October to a healthy baby girl.

Sukin is not sure what she would have done had she found out five years ago that she was carrying a child who might have Angelman's syndrome. But she is sure she would have wanted to know. "We have a responsibility to share whatever knowledge is available," Sukin says. "The majority of people will have a healthy child. But when you're the statistic, your one kid is the most important thing." ■

Erika Check is Nature's Washington biomedical correspondent.



For Anca Segall, taking a full prenatal screening test was important for her pregnancy.

A. SEGALL & F. ROHWER

1. de Vries, B. B. *et al. Am. J. Hum. Genet.* **77**, 606–616 (2005).
2. Shaw-Smith, C. *et al. J. Med. Genet.* **41**, 241–248 (2004).
3. Bejjani, B. A. *et al. Am. J. Med. Genet. A* **134**, 259–267 (2005).
4. Check, E. *Nature* **437**, 1084–1086 (2005).
5. Rickman, L. *et al. J. Med. Genet.* doi:10.1136/jmg.2005.037648 (2005).
6. Le Caignec, C. *et al. J. Med. Genet.* **42**, 121–128 (2005).

BUSINESS

Internet star shoots for a rocket revolution

A newcomer to the space arena claims that his start-up launch company can buck the trend of commercial failures. Tony Reichhardt reports.

For the past 20 years, most start-up rocket companies have followed the same sad trajectory. They produce an artist's concept of an innovative vehicle while promising to cut launch costs by a factor of ten. Scrounge for money, mostly without success. Badmouth NASA and established rocket manufacturers such as Boeing. Fail, usually before reaching the launch pad. And disappear.

The names may be different — AMROC, Conestoga, Rotary Rocket, Beal Aerospace — but the stories are essentially the same. And their combined impact on the economics of spaceflight has been zero. It still costs tens of millions of dollars to place even a modest satellite in orbit.

Enter Elon Musk, the latest comer with long-shot dreams of revolutionizing the launch business. A 34-year-old South African Internet tycoon who made his fortune inventing the PayPal online payment service, Musk has been in the rocket game since 2002. His Falcon 1 launch vehicle — which he hopes will become the first privately funded US rocket to reach orbit since Pegasus in 1990 — is set to launch from the Kwajalein atoll in the Pacific later this month, after a 26 November attempt was aborted because of a faulty fuel-valve setting.

If his lofty ambitions materialize, Musk says, this launch will mark only the beginning. He plans a whole family of rockets, from the Pegasus-class Falcon 1 (list price US\$6.7 million) to the Falcon 9-S9, which would compete with the biggest Delta and Atlas rockets in existence and undercut their \$150-million-plus pricing by half.

Big talk. And unlikely to happen, according to veterans of the launch business. Wolfgang Demisch, a Wall Street financial consultant who has tracked aerospace investments for more than 30 years, testified in April to Congress on the prospects for commercial



Rocket man: Elon Musk (left) has put \$100 million of his own money into the production of the Falcon 1 vehicle (above), due for launch this month.

space activity — with Musk sitting alongside him at the witness table. “Regrettably, I am unaware of any credible proposal to achieve the desired substantial cost reductions,” he said.

Why, after all these years, is space launch still so expensive? Because it is really, really difficult, answers Antonio Elias, a vice-president at Orbital Sciences who led the design team for Pegasus. “You’re essentially riding on the edge of catastrophe all the time,” he says. Sending rockets into space is far riskier than other

high-tech ventures such as the airline business, and operates with much thinner safety margins. Most practical improvements in the basic technology, whether in choice of fuels or structural materials, were thought of long ago. And there are no efficiencies to be gained from mass production. “We’re not building 20,000 units,” says Elias. “It is still a handmade item.”

As a result, launch costs have remained stubbornly high. In the decade from 1990 to 2000, Western-built small rockets cost an average of \$8,445 per pound delivered to orbit, according to a 2002 study by aerospace business analyst Futron. Pegasus, the cheapest existing US launch system, with a 1,000-pound capacity, currently goes for more than \$15 million a ride. Russian and Chinese rocket-builders, with far lower labour costs, could easily beat that, but have to keep their rates high enough to avoid charges of price dumping by US and European competitors.

Money talks

Into this tough and idiosyncratic market charges Musk, who has already spent \$100 million of his own cash — his net worth is estimated at three times that — to bankroll the development costs of Falcon 1 at his California-based company, SpaceX. Experts say that, so far, he has got good value for his money.

Elias, for one, is “very impressed” that the company managed to develop a liquid-fuelled rocket from scratch in just a few years. Falcon’s kerosene-fuelled Merlin engine has performed well in ground tests, and now, according to the company, runs trouble-free. SpaceX has strived for simplicity in building Falcon. The two-stage rocket has only one engine per stage, and the separation of one stage from another is designed to be as simple as possible. Futron, in a November 2004 analysis of rocket reliability, gave the Falcon, an unproven vehicle, a rating better than or equal to any existing US launch system.

But one perfect launch, or several in a row, won’t ensure long-term financial success, let alone a dramatic breakthrough in pricing. A little rocket such as Falcon 1 can be operated by a small, dedicated team, but experts say Musk will have difficulty keeping costs down as he scales up to bigger vehicles. “They’ll have more metal and they’ll have more fuel. Then they’ll need a bigger building to hold all that stuff in,” says Leon McKinney, an aerospace consultant based in St Louis, Missouri, and a member of the Space Transportation Technical Committee for the American Institute of Aeronautics and Astronautics.

Not to mention the inevitable growth in the organization. The launch business is

C. THOMPSON/SPACEX

D. KIM/SPACEX

notoriously prone to government supervision and delay. Even if Musk runs a lean manufacturing shop, he will face regulatory requirements and other obstacles imposed by government agencies and will need more people to deal with the red tape.

Limited market

Then there is the competition. Musk has his eye on a key NASA contract, details of which are expected to be announced this month, to provide a commercial cargo delivery service to the International Space Station. It could be worth up to \$200 million over several years, and Musk says he will pursue it vigorously. But so will every other rocket company.

What's more, the overall market for rocket launches remains limited. Forecasters predict only 15 to 20 commercial launches a year, worldwide, for the foreseeable future. And even if space tourists continue to pay \$20 million for a ride into orbit, as they have done in the past to visit the space station, the number

of tickets is likely to be limited.

"One of my core beliefs is that we should be a space-faring civilization."

— Elon Musk

Musk has heard all of this, and hopes to shake things up anyway — which, given his wealth, isn't entirely out of the

question. He is also good with politicians and the press, managing to come across as both guileless and smart. He had the sense to hire experienced rocket engineers to build Falcon. And he's not afraid to go to court when he thinks the established rocket manufacturers are engaged in unfair trade practices. In October, he sued Boeing and Lockheed Martin in an attempt to block a proposed merger of their Atlas and Delta rocket programmes that would better position them to win government launch contracts. "Mr Musk knows how to take care of himself," says McKinney.

The real test will be his staying power. If the first Falcon blows up, Musk has said he will try at least two more times before giving up. If it's successful, the second launch will be in March. If things are still going well at that point, he plans to court outside investors to raise a further \$100 million to build his larger rockets. If none steps forward, he adds, he will put the money up himself.

Musk says he got into rocketry because "one of my core beliefs is that we should be a space-faring civilization". Elias of Orbital Sciences sounds sincere when he wishes him luck. But he says: "Nothing in the laws of physics and very little in the laws of economics gives me hope."

IN BRIEF

VACCINE BUYOUT In the latest evidence of consolidation in the vaccine industry, the Dutch biotechnology firm Crucell agreed last week to buy the Swiss vaccines company Berna Biotech for €381 million (US\$449 million). The all-share offer, expected to be launched in mid-December, will value Berna shares at a premium of 27% on their closing price on 30 November, the day before the deal was made public. The acquisition creates a company big enough to compete with large vaccine units such as those at Sanofi-Aventis, GlaxoSmithKline and Chiron, which is being acquired by Novartis.

GOING GREEN BP, Europe's leading oil company, plans to double its investment in alternative energy, spending up to US\$8 billion over ten years on power generation from solar, wind, hydrogen and combined-cycle-gas-turbine sources. The company last week launched a new unit, BP Alternative Energy, to manage the programme. It predicted that the business could generate revenues of \$6 billion annually within a decade. The new business unit will employ some 2,500 people globally.

LUCRATIVE LICENCE An Alabama biotechnology firm has licensed a drug in early clinical trials to Roche in a deal potentially worth more than half a billion dollars. BioCryst Pharmaceuticals of Birmingham said last week that the Swiss drug-maker will pay it \$25 million up front for worldwide rights to BCX-4208, a drug intended to prevent transplant rejection and to treat patients with autoimmune diseases. Payments by Roche for future milestones could reach \$530 million. BioCryst shares surged 37% after the deal was announced.

MARKET WATCH



This week Wood Mackenzie, an Edinburgh-based research and consulting firm, reviews recent trends in biotechnology stocks.

The Nasdaq biotechnology index fell sharply in early October, losing 9% of its value following a three-month rally. The decline was influenced by industry-specific events, but also by broad market trends, with the fall being mirrored by other, more general indices. The index recovered to end the period just 0.7% down, driven by licensing deals as well as strong earnings for the year's third quarter reported in November. All told, the index is up 3.6% on the year to date.

Clinical-trials failures contributed to the October fall: Human Genome Sciences of Rockville, Maryland, saw its shares plunge 45% after clinical results showed that its antibody drug candidate LymphoStat-B was ineffective in treating lupus. The drug later partially redeemed the company's share value with encouraging clinical-trials results for rheumatoid arthritis, limiting the loss

over the period to 35%. NABI Pharmaceuticals of Boca Raton, Florida, rode out the early fall only to take a 74% nosedive in share value in early November after poor clinical-trials results halted further development of its StaphVax vaccine to prevent dangerous bacterial infections.

The market recovery was supported by deal-making activity. Early last month, Pain Therapeutics of South San Francisco entered a strategic alliance valued at more than \$400 million with King Pharmaceuticals of Bristol, Tennessee, to develop an addiction-resistant narcotic painkiller. And in late November, Incyte Pharmaceuticals of Wilmington, Delaware, signed a broad research and development agreement worth up to \$800 million with drug giant Pfizer. Despite the November surge, the Nasdaq biotechnology index has significantly underperformed compared with wider market indices in the past month.

www.woodmac.com

Supplementary data need to be kept in public repositories

SIR — The reality of the genomics age is that there are many very large data sets that are most usefully saved and manipulated in electronic form. Many journals add online ‘supplementary material’ to articles as a service to authors wishing to publish volumes of such data that cannot be accommodated within the body of an article.

Supplementary-material collections maintained by publishers serve as archival repositories directly connected with the peer-reviewed scientific literature, often competing with or substituting for the deposition of data in public repositories.

To assess the use of these, we investigated supplementary-data archives for gene-expression profiling data, a widely used experimental protocol for which international standards for data representation have been developed.

We anticipated that such archives might be a useful source of data. But to our dismay, it was impossible to systematically analyse our sample, taken from 10,128 papers in 139 journals. No standards for organizing supplementary-data collections have been adopted either across journals or even for supplementary-data collections associated with articles in the same journal.

Data are represented in an enormous range of different file formats, from raw data files (such as Affymetrix.cel files) to spreadsheets (xls file extensions), documents (doc and pdf) and text files (txt and cvs). Within documents there are no standards for data organization: different documents provide different numbers of columns, contain both differential and absolute expression values, and often have few details about the signal processing applied to obtain data. We also encountered a significant number of typographic errors in gene names, database accession numbers and data-set identifiers.

There are public repositories for gene-expression profile data (Stanford MicroArray Database, the US National Center for Biotechnology Information Gene Expression Omnibus and the ArrayExpress repository at the European Bioinformatics Institute). We compared the accessibility of gene-expression profile data in public repositories with accessibility of data in supplementary-data archives. The public repositories provide numerous search and retrieval tools, including unique accession numbers and the ability to search by specimen, platform and profile data. Publishers’ supplementary-materials archives provided none of these features. As a result, relevant data are far harder to locate than in public repositories.

These findings are not limited to

gene-expression data. Even within the same journal, there is no consistency in reporting or format among bioinformatics resources. File extensions for documents, figures and movies include xls, doc, eps, jpg, tif, gif, pdf, ppt, qt, asf, wma and wmv. They may or may not include long lists of links, be compressed into zip files or offer the option of including the supplementary material as part of the downloadable document containing the printed version of the article.

Supplementary data often represent the raw experimental values and are especially important for researchers in the same field. Among the advantages of storing these data in public repositories are the integration of information with the community knowledge resources and the ability to track and maintain computer-readable associations between data sets.

On the basis of our analysis, we recommend that scientific journals adopt a policy, similar to *Nature*’s (see www.nature.com/nature/authors/policy/index.html#a7.2), of requiring that authors submit data to public repositories, if relevant repositories exist, and that the journal version should contain accession numbers, URLs and other appropriate specific indicators to the data source in the repositories.

Journals’ supplementary-data archives should be restricted to idiosyncratic and nonstandard data types for which no public repository exists. Only then can community standards emerge.

Carlos Santos*, **Judith Blake†**, **David J. States*‡**
 *Bioinformatics Program, University of Michigan, Ann Arbor, Michigan, USA

†The Jackson Laboratory, Bar Harbor, Maine, USA

‡Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA

Turkish science needs more than membership of the EU

SIR — Your Editorial “Turkey’s evolution” (*Nature* **438**, 1–2; 2005), about the country’s efforts to join the European Union (EU), states that “the opening of negotiations for EU membership offers the best hope for the continuing development of science in Turkey”. This view is common in Europe, but I believe the assumptions behind it lack solid support.

First, you assume that EU policies adopted by Turkey during membership negotiations will lead to more economic investment in Turkish science. Such investment is needed if Turkey is to close the gap with more developed countries. But the increase in the science budget, to US\$300 million in a country of 70 million, is inadequate. The €250 million (US\$292 million) that Turkey contributed towards the EU’s Sixth Framework programme is not expected to be recouped. And even though policies prescribed by

the International Monetary Fund (IMF) have reduced investment in the country’s educational infrastructure (E. Voydova and E. Yeldan *Comp. Econ. Stud.* **47**, 41–79; 2005), keeping to an IMF programme is a condition for Turkey’s acceptance into the EU.

Second, although international scientific collaboration is crucial for scientific development in any country, the extent to which knowledge sharing and cooperation depends upon international economic and political relations is less clear. Some countries, such as Cuba, India and China, have achieved scientific progress in relatively independent economic or political circumstances. Political and cultural relations among countries at dissimilar levels of development might even impede progress on the weaker side — for example, through a ‘brain drain’ effect.

Last, I fear that entrusting all hope of development to the ambiguous political process of EU membership may undermine Turkey’s existing — albeit weak — resolution to advance science.

The country needs a firm political resolution to implement long-term public investments in education and science, regardless of EU membership negotiations.

Mehmet Somel

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany

Flu virus will not be sent in the regular US mail

SIR — The headline and photographs of your News story “Deadly flu virus can be sent through the mail” (*Nature* **438**, 134–135; 2005) are misleading with respect to the policy of the Centers for Disease Control and Prevention (CDC) regarding the transfer and use of the 1918 pandemic influenza virus. They could give the erroneous impression that the virus will be made widely available and sent through the regular US mail.

The CDC has not yet received any requests to work with the 1918 virus at a non-CDC facility and I have made it clear we currently have no plans to send the virus anywhere. Any requests we do receive will be considered on a case-by-case basis, taking into account scientific merit, biosafety and biosecurity concerns, as well as any additional standards deemed appropriate for this particular virus.

The CDC is the only agency that currently possesses this virus, and we have a special responsibility to balance the importance of scientific progress and collaboration with the moral and scientific imperatives of biosafety and biosecurity.

Julie Louise Gerberding

Centers for Disease Control and Prevention, 1600 Clifton Road Northeast, Atlanta, Georgia 30333, USA

BOOKS & ARTS

Physics ain't what it used to be

Science is venturing into areas where experimental verification simply isn't possible.

The Cosmic Landscape: String Theory and the Illusion of Intelligent Design

by Leonard Susskind

Little Brown: 2005. 416 pp. \$24.95

George Ellis

Once upon a time, physics dealt with tangible objects — if you couldn't weigh them or smash them together, at least you could observe them. As times changed, physicists started to deal with more ethereal things: electromagnetic fields and space-time metrics, for example. You couldn't see them but you could measure their influence on particle trajectories and so justifiably claim evidence of their existence. Nowadays things have changed. A phalanx of heavyweight physicists and cosmologists are claiming to prove the existence of other expanding universe domains even though there is no chance of observing them, nor any possibility of testing their supposed nature except in the most tenuous, indirect way.

How can this be a scientific proposal, when the core of science is testing theories against the evidence? In *The Cosmic Landscape*, Leonard Susskind argues that we should accept the reality of such universe domains on the basis of two theoretical elements that, taken together, could provide a solution to two major scientific conundrums. The first puzzle is the anthropic issue: the "apparent miracles of physics and cosmology" that make our existence possible. Many aspects of both physics and cosmology seem to be fine-tuned in such a way as to allow chemistry to function, planets to exist, and life to come into being. If they were substantially different, no life at all, and so no processes of darwinian evolution, would have occurred. Which particular aspect of this fine-tuning seems the most significant depends on one's discipline.

Susskind, a particle physicist, thinks the most important is the issue of the cosmological constant, relating to a universal repulsive force that acts on all matter. But this leads to the second conundrum: simple estimates suggest that this constant should be 120 orders of magnitude larger than recently observed. This is a major crisis for quantum field theory, which underlies these estimates. The link to the anthropic question is that if the constant were only twice as large, there would be no galaxies, stars, planets or life. The observed very small value of this constant, although contrary to

our present theory of the quantum vacuum, is a necessary condition for our existence.

The first part of the proposed solution is the idea of a 'multiverse' — the existence of a huge number of 'pocket universes', like the vast expanding Universe domain we see around us, that are part of a much larger physical existence. These are supposed to arise through inflation, a process of extremely short-lived, very rapidly accelerating expansion that preceded the hot Big Bang era in the early Universe. 'Chaotic inflation' occurs if inflation is still occurring in distant domains around us today, forming overall a fractal-like structure of inflating domains and pocket universes.

The second part of the solution is the landscape of possibilities, a recent discovery in string theory, which is itself a proposed theory of fundamental physics that unites gravity with quantum physics. It has been suggested that the 'vacuum' of string theory is a structure of immensely complex possibilities, with each possible vacuum resulting in a different kind of local physics; for example, all possible values of the cosmological constant will occur in the different vacua of string theory. If we suppose that the pocket universes of chaotic inflation correspond to different vacua, then all possible kinds of local physics occur at different loca-

tions somewhere in the multiverse. If enough combinations of possibilities are realized in this way, then the incredibly special conditions for life to exist will inevitably occur somewhere in the multiverse. The apparent design of conditions favourable to life in our own universe domain can therefore be explained in a naturalistic way.

This is an intriguing picture that unites quite disparate elements of physics and cosmology in a synthesis that is satisfying in many ways. But the question here is whether it is a scientific proposal, as there is no chance whatsoever of observationally verifying its main prediction, the existence of numerous other expanding universe domains beyond our visual horizon. We might hope to base our prediction that the multiverse exists on the fact it is an inevitable outcome of well established physics, but the physics underlying the proposal is hypothetical, rather than established. String theory is neither well defined nor experimentally proven, despite the energy and enthusiasm of its proponents, and there are alternative theories. The inflation field has not been uniquely identified in physical terms, much less shown to have the properties supposed in chaotic inflation.

We might hope to detect the multiverse

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

PETER ARNOLD/ALAMY

You gotta have faith: the idea of countless 'pocket universes' cannot be tested, so is it science?

indirectly by observing the remnants of the physical processes that underlie its existence; for example, the low value of the cosmological constant today could be such a hint. The problem here is that a multiverse proposal cannot in general be disproved this way, because if all possibilities exist somewhere in the multiverse, as some claim, then it can explain any observations, whatever they are. For example, no observations of anisotropy in the cosmic background radiation can disprove the multiverse hypothesis because all possible anisotropies will be generated in the different expanding universe domains; you just have to live in the right one.

The particular multiverse version proposed by Susskind, however, has the great virtue of being testable in one respect. It is supposed to have started out by quantum tunnelling, resulting in a spatially homogenous and isotropic universe with negative spatial curvature, and hence with a total density parameter $\Omega_0 < 1$. The best observationally determined value for this parameter, taking all the data into account, is $\Omega_0 = 1.02 \pm 0.02$. Taken at face value, this seems to contradict the proposed theory. But given the statistical uncertainties, the observations do not definitively exclude $\Omega_0 < 1$, so the theory survives; nevertheless, the observed value should be taken seriously in this era of 'precision cosmology'. These data are not discussed in the book — a symptom of some present-day cosmology, where faith in theory tends to trump evidence. Presumably the hope is that this observational result will go away as more evidence is collected.

The Cosmic Landscape is extremely well written, provides an excellent non-technical overview of the relevant physics, and tackles important questions in a lively way. However, it confuses the event horizon in the expanding universe with particle and visual horizons. In addition, like many multiverse writings, it uses the concept of infinity with gay abandon, when there is good reason — as pointed out by mathematician David Hilbert — to claim that it is not a good physical concept. The book also tries to justify the multiverse idea in terms of the 'many worlds' interpretation of quantum theory — an unproven and totally profligate viewpoint that many find difficult to take seriously.

As a philosophical proposal, the multiverse idea is interesting and has considerable merit. The challenge facing cosmologists now is how to put on a sound basis the attempts to push science beyond the boundary where verification is possible — and what label to attach to the resultant theories. Physicists indulging in this kind of speculation sometimes denigrate philosophers of science, but they themselves do not yet have rigorous criteria to offer for proof of physical existence. This is what is needed to make this area solid science, rather than speculation. Until then, the multiverse situation seems to fit St Paul's description: "Faith is the substance of things hoped for, the

evidence of things not seen." In this case, it is faith that enormous extrapolations from tested physics are correct; hope that correct hints as to the way things really are have been identified from all the possibilities, and that the present marginal evidence to the contrary will go

away. This book gives a great overview of this important terrain, as seen from an enthusiast's viewpoint. ■

George Ellis is in the Department of Mathematics, University of Cape Town, Cape Town 7701, South Africa.

Pet project

The Dog and Its Genome

edited by Elaine A. Ostrander, Urs Giger & Kerstin Lindblad-Toh

Cold Spring Harbor Laboratory Press: 2005. 584 pp. \$135, £80

Stephen J. O'Brien

Genome technology has found its way into the living room with the completion of the whole-genome sequence of the domestic dog *Canis familiaris*, from a female boxer called Tasha. Finished just a year after its initiation in 2003, the remarkably complete sequence (representing an estimated 99% of the dog's 2.4 billion base pairs) achieves 7.5-fold coverage of the genome and is a major advance over the 1.5-fold sequence of a poodle published by Celera in 2003. The dog is now a front-line model for the discovery of disease genes, for gene annotation, and for probing the evolutionary roots of our mammalian origins. *The Dog and Its Genome*, edited by Elaine Ostrander, Urs Giger and Kerstin Lindblad-Toh, celebrates the completion of the dog sequence with 26 chapters on the genomic biology of man's best friend.

The book should appeal to dog fanciers, to genome biologists who wonder about the sequence's applications, and to students of comparative genomics. It presents well written and concise discussions of the history of dog breeds — there are generally estimated to be between 350 and 1,000, of which the American Kennel Club recognizes about 150 that do not exchange genes. As many as 20 breeds were developed by 1750, increasing to 76 by 1905. Yet the domestication of dogs can be traced back 14,000 years on the evidence of archaeological remains, maybe even 40,000 years based on molecular comparisons with wolves. Clearly, dogs are the oldest domesticated species, as detailed in two of the book's chapters, and the phylogenetic ancestry of dog breeds is described in three chapters.

Years from now, as dog genomics matures, this volume will be remembered as the starting point, with vivid pieces on the vast phenotypic variation described for dogs. The latest interpretation of dog genome status is presented for experts and aficionados alike. The remarkable history of inbreeding has led to a mosaic genome of alternating homozygous and heterozygous/polymorphic segments specific for each breed; these are particularly useful for linkage disequilibrium-based association mapping of complex or multifactorial traits.



Boxer tricks: Tasha's genome will help researchers to understand human genetic diseases.

And dogs certainly have complex traits, notably the vast morphologic variation found in dog breeds as disparate as the chihuahua and the great dane. Dogs also have hard-wired behavioural acumen that allow them to herd livestock, locate missing persons and even sniff out human cancers at early stages. And of course they are loving companions like no other animals.

Generations of veterinary clinicians have identified nearly 500 human hereditary disease homologues in dogs, nearly all breed-specific; the 50 reviewed here have a confirmed genetic basis. Several have been treated successfully with futuristic gene-therapy protocols that should whet the appetite of the medical community. The book describes a cancer registry that documents the incidence and pathologies of a dozen neoplasms that account for 23% of deaths in the 65 million pet dogs in the United States. The challenge now will be to use the genome to detail the genetic bases of behaviours, morphological breed distinctiveness and the disposition of breed-specific cancers.

Researchers already have 'bibles' that define gene-based phenotypes suitable for interrogation by mouse, rat, fruitfly and human genetics. *The Dog and Its Genome* does the same for the canine genomics community. It should be consumed by researchers and their students quickly before forthcoming advances render it dated on their bookshelves. ■

Stephen J. O'Brien is in the Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland 21702-1201, USA.

NHGR/MIT



On top of the world

Nanga Parbat in Pakistan, the ninth highest mountain in the world, is one of 90 or so mountains included in *Mountains From Space: Peaks and Ranges of the Seven Continents*. This striking book, which contains photos taken from space, is published by the German Aerospace Centre (DLR) and Harry N. Abrams, and is available in German and English editions.

Taking its name from the Sanskrit for 'sacred mountain', Nanga Parbat is 8,125 metres high and has cost many mountaineers their lives, including Günther Messner, whose brother and climbing partner Reinhold came up with the idea of the book. Reinhold Messner is among several authors who provide essays from their perspectives as mountaineers or scientists.

This picture was taken by the SPOT-5 Earth-observation satellite. **A.A.**

EXHIBITION

A close look at Darwin

Darwin

American Museum of Natural History, New York, until 29 May 2006.
www.amnh.org/exhibitions/darwin

Alan Packer

The American Museum of Natural History in New York bills its new exhibition, *Darwin*, as the most in-depth ever mounted on Charles Darwin's life and thought. It's also well timed, coming as it does in the midst of litigation over 'intelligent design' in Dover, Pennsylvania, and in the run-up to the bicentennial of Darwin's birth in 2009. All that aside, *Darwin* is splendid: evolutionary biologist Niles Eldredge's exhibition takes us on a fascinating tour through the life of a great thinker, in what is a superb example of the curator's art.

Visitors are greeted at the entrance by a live, and somehow mesmerizing, giant tortoise from the Galapagos (it can also be viewed remotely via webcam at the exhibition's website). They then encounter Darwin's magnifying glass. This serves as an iconic image — throughout the exhibition, a magnifying glass is positioned to allow the viewer a closer look at a specimen, symbolizing the overall theme of Darwin's lifelong devotion to close observation of nature. His theoretical conclusions, which are well explained in the exhibition, rest on a mountain of evidence that he saw with his own eyes. That evidence is presented here in abundance.

There are many noteworthy items on display, including original correspondence, specimens from Darwin's own collection, original notebooks, and pressed plants from the voyage of the *Beagle*. The writing box that belonged



Among other exhibits, visitors to *Darwin* can see a cast of a glyptodont skeleton found by Darwin.

to his daughter Annie is included here (she died at the age of 10). So too is an amusingly exhaustive questionnaire that Darwin sent to 'gentleman farmers', enquiring about their experiences with artificial breeding. The atmosphere is congenial: sounds of ocean life are heard in the section devoted to Darwin's voyage on the *Beagle*; there is a 'condensed-time' video of the 'Sandwalk' footpath around Darwin's home and workplace, Down House in Kent, UK; and in the exit room, a voiceover of the final words from Darwin's book *On The Origin of Species* ushers you through a collection of orchids.

"Believing is easy, and knowing is hard, and it's knowing that matters most," wrote Neil

Patterson in his introduction to cell biologist Christian de Duve's recent book *Singularities* (Cambridge University Press, 2005). Darwin removed the nebulous idea of belief from the discussion. In explaining what we know about the theory of evolution and its originator, given the limitations of what an exhibition can convey, *Darwin* could hardly be bettered.

When the *Darwin* exhibition closes at the American Museum of Natural History in New York on 29 May 2006, it will travel to the Museum of Science in Boston, The Field Museum in Chicago, the Royal Ontario Museum in Toronto and the Natural History Museum in London.

Alan Packer is senior editor at *Nature Genetics*. ■

The message of the quantum

Einstein challenged physics to describe “the real factual situation”. But an understanding of the very concepts that he criticized a century ago may provide the best clues yet about reality ‘out there’.

Anton Zeilinger

In the first of his papers from 1905, his *annus mirabilis*, Einstein proposed the idea of particles of light, later called photons. From this paper, a very realistic picture of light particles emerged, as being much like the particles in an ideal gas. But the paper also contained the seeds of Einstein's later criticisms of quantum mechanics. As he described in his *Autobiographical Notes*, Einstein challenged physics, including the concepts of quantum mechanics, to describe “the real factual situation”, or, in other words, what is out there.

The concepts that Einstein criticized were randomness, entanglement and complementarity. These have become the core principles of newly emerging quantum information technologies: quantum computation, quantum teleportation and quantum cryptography. But although we may have realized that Einstein was wrong about these concepts, have we today understood the message of the quantum?

The discovery that individual events are irreducibly random is probably one of the most significant findings of the twentieth century. Before this, one could find comfort in the assumption that random events only seem random because of our ignorance. For example, although the brownian motion of a particle appears random, it can still be causally described if we know enough about the motions of the particles surrounding it. Thus, as Werner Heisenberg put it, this kind of randomness, of a classical event, is subjective.

But for the individual event in quantum physics, not only do we not know the cause, there is no cause. The instant when a radioactive atom decays, or the path taken by a photon behind a half-silvered beam-splitter are objectively random. There is nothing in the Universe that determines the way an individual event will happen. Since individual events may very well have macroscopic consequences, including a specific mutation in our genetic code, the Universe is fundamentally unpredictable and open, not causally closed.

Most striking is the case of entanglement, which Einstein called “spooky”, as it implies that the act of measuring a property of one particle can instantaneously change the state of another particle no matter how far

apart the two are. Distances over which this phenomenon have been verified experimentally are in the order of 100 kilometres. How is it possible that two events, each one objectively random, are always perfectly correlated?

John Bell showed that the quantum predictions for entanglement are in conflict with local realism. From that ‘natural’ point of view any property we observe is (a) evidence of elements of reality out there and

of heavenly bodies. I suggest that in a similar way, the distinction between reality and our knowledge of reality, between reality and information, cannot be made. There is no way to refer to reality without using the information we have about it.

Maybe this suggests that reality and information are two sides of the same coin, that they are in a deep sense indistinguishable. If that is true, then what can be said in a given situation must, in some way, define, or at least put serious limitations on what can exist.

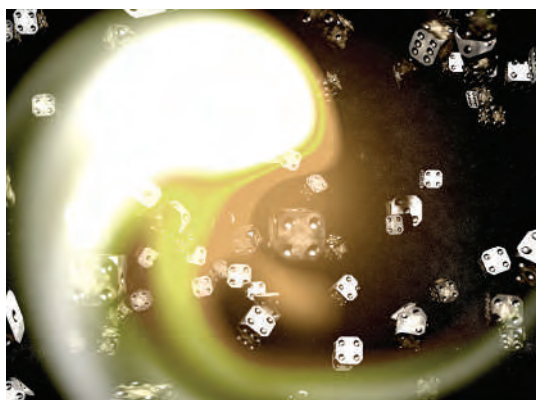
These ideas can be brought to fruition through understanding the three concepts criticized by Einstein. It is natural to assume that the information represented by a quantum system scales with its size. The randomness of the individual event is then a direct consequence of the fact that not enough information is available to pre-define the outcomes of all possible measurements. The same holds for complementarity, implying that the information available only suffices to define the outcomes of one of a number of mutually complementary measurements. Finally, entanglement is the observation that the finite information available to characterize two (or more) systems can either be used to define the properties of the individual systems, as in classical physics, or to define the results of joint observations of both or all systems together.

So the experimentalist, by choosing the apparatus, can define which quality of a number of possibilities will become reality in the measurement. But the individual measurement result remains objectively random because of the finiteness of information. I suggest that this randomness of the individual event is the strongest indication we have of a reality ‘out there’ existing independently of us. Maybe Einstein would have liked this idea after all. ■

Anton Zeilinger is at the Institute for Experimental Physics, University of Vienna, Boltzmanngasse 5, A-1090 Vienna, Austria, and at the Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria.

FURTHER READING

Zeilinger, A., Weihs, G., Jennewein, T. & Aspelmeyer, M. *Nature* **433**, 230–238 (2005).
Schilpp, P. A. *Albert Einstein: Philosopher-Scientist* (Open Court Publishing, Library of Living Philosophers, Peru, Illinois, 1949).



Chancing it: the Universe is fundamentally unpredictable.

(b) independent of any actions taken at distant locations simultaneously with the measurement. Most physicists view the experimental confirmation of the quantum predictions as evidence for nonlocality. But I think that the concept of reality itself is at stake, a view that is supported by the Kochen–Specker paradox. This observes that even for single particles it is not always possible to assign definite measurement outcomes, independently of and prior to the selection of specific measurement apparatus in the specific experiment.

A criticism of realism also emerges from the notion of complementarity. It is not just that we are unable to measure two complementary quantities of a particle, such as its position and momentum, at the same time. Rather, the assumption that a particle possesses both position and momentum, before the measurement is made, is wrong. Our choice of measurement apparatus decides which of these quantities can become reality in the experiment.

So, what is the message of the quantum? I suggest we look at the situation from a new angle. We have learned in the history of physics that it is important not to make distinctions that have no basis — such as the pre-newtonian distinction between the laws on Earth and those that govern the motion

C. DARKIN

ESSAY

NEWS & VIEWS



C. COLLINS/CORBIS

GENOMICS

The dog has its day

Hans Ellegren

Domestication and selective breeding have transformed wolves into the diversity of dogs we see today. The sequence of the genome of one breed adds to our understanding of mammalian biology and genome evolution.

Dogs have a special place in our society. Man's best friend is not just a valuable hunting partner, guard and herd manager — most of the world's estimated 400 million dogs¹ are pets. Dogs were the first animals to be domesticated (at least 15,000 years ago)^{2–4}. They all originate from a single and relatively homogeneous species — the wolf — but modern breeds display an extraordinary diversity of traits (or phenotypes). The hundreds of years of careful inbreeding to produce the many kinds of dog have delivered a geneticist's dream model of human genetic disease (Box 1, overleaf). But to unlock the full potential of this model, we need to understand the genetic basis for the unprecedented diversity and how it has evolved⁵. The high-quality draft sequence of the dog genome described on page 803 of this issue⁶ is a good starting-point for that research.

Lindblad-Toh and colleagues⁶ invited breed clubs and veterinary schools to suggest an individual dog suitable for genome sequencing. The idea was to identify a highly inbred dog; this was based on the thinking that the animal's genetic homogeneity would simplify the gigantic jigsaw puzzle of assembling millions of sequence reads into a genome sequence. After testing certain genetic markers in a host of dogs, the sequencers settled on a female boxer called Tasha (so there is no Y chromosome in the current sequence).

The assembled sequence from Tasha's DNA spans 2.4×10^9 base pairs (Gbp), which corresponds to an estimated 99% coverage of the canine genome (excluding highly repetitive regions). So, although dogs have 39 pairs of chromosomes (compared with 23 pairs in humans), their genome contains almost 0.5 Gbp less DNA than ours. The difference can be explained mainly by the existence of fewer repetitive elements in the dog lineage, and to some extent by deletion of sequences that were present in an early common mammalian ancestor. Dogs seem to have fewer genes than humans, but the actual numbers might be a bit out for both genomes because identifying genes across whole genomes continues to be a difficult task⁷.

The current work is not the first canine genome project. Sequencing of a male poodle (at a lower sequence coverage) recently characterized about 75% of its genome, although with much of the assembled sequence interleaved with gaps of undefined length⁸. However, by comparing it with the boxer genome, the poodle sequence is a useful tool for identifying genetic variants — single nucleotide polymorphisms (SNPs) — in dog populations. Augmented with SNPs identified in the boxer and by limited sequencing of many other dog breeds, 2.5 million variable sites have now been discovered⁶. Comparisons of

the different breeds show that there is an average of around 1 SNP per 1,000 base pairs — a similar value to that in human populations.

The SNP data give several evolutionary insights. For instance, analysis of DNA from mitochondria (cellular organelles that have their own genome) has suggested that domestication is associated with a narrow genetic bottleneck where only a few wild ancestors contributed to the domestic gene pool⁹. However, the large genetic diversity seen among dogs is at odds with this hypothesis, and work on other domestic animals shows that they, too, have high levels of variability in their nuclear genes. This implies that, in many cases, back-crosses with wild relatives introduced additional genetic diversity into domesticated animals well after domestication began¹⁰. The genetic traces of such interbreeding may not be picked up by studies of mitochondrial DNA if the back-crossing occurred mainly between wild males and domestic females, because mitochondrial DNA is inherited only from mothers¹¹.

The physical positions of the genetic variations within and among breeds create patterns in the genome that give a more detailed perspective on domestication and breed formation. Within breeds, most chromosomes are mosaics of alternating regions of homogeneous sequences — reflecting the recent

Box 1 | From wolf to dog to disease model

The wild ancestor of dogs, the grey wolf, belongs to a large group of mammals called the Carnivora, which includes cats, bears and seals. Roughly 40 million years ago, a family of dog-like carnivores (Canidae) evolved, and about 15 million years ago they diverged into foxes, wolves, jackals and others. A phylogenetic analysis⁶ shows that the coyote is the closest living relative of the grey wolf (the two species had a common ancestor one million to two million years ago), followed by, in order of genetic distance, the golden jackal, Ethiopian wolf, dhole and African wild dog. Ancient dog remains from Alaska and Latin America indicate that native American dogs originated from dogs domesticated in the Old World⁴. These

dogs must have accompanied late Pleistocene humans across the Bering Straits, which means they were domesticated at least 15,000 years ago, probably in southeast Asia³.

Modern dog breeds have subsequently been generated by selecting for existing traits among the wild ancestors — a prime example of evolution by selection. The extraordinary variation in shape, size, behaviour and physiology of the breeds makes the dog a unique genetic model; each pure breed is an inbred, isolated genetic population, with simplified genetic structures that can be linked to their physical traits.

Several hundred genetic disorders shared between dogs and humans have been reported,

many of which are found in just one or a few breeds. For instance, narcolepsy is seen largely in doberman pinschers, and a hereditary kidney cancer occurs only in German shepherd dogs; the genes underlying both diseases have been identified in dogs. Examples of genetic diseases common to several breeds include blindness, allergy and epilepsy. Using dogs as a model for human genetic disease can not only identify causative genes and aid the development of treatments, but can also provide information on the character of disease-causing mutations. For example, the expansion or insertion of repetitive elements in genes has recently been shown to cause disease in both humans and dogs^{16,17}. **H.E.**

common ancestry shared by individual dogs of the same breed — and heterogeneous sequences^{6,12,13}. Mathematical simulations can be used to model the way in which population history might be expected to affect genetic diversity and its structural patterns. The model that best fits the observed pattern of SNPs is one that assumes an ancient bottleneck some 9,000 generations ago (domestication), followed by breed-specific bottlenecks 30–90 generations ago (breed formation). However, if repeated back-crossing has occurred, this model would have to be revised.

The dog adds to a growing list of vertebrate species that have had their genome sequenced¹⁴. A comparative analysis of the human, mouse and dog by Lindblad-Toh *et al.*⁶ showed that about 5% of the human genome is being maintained by natural selection — suggesting that it has some essential function. Almost all of this sequence is also present in the dog genome. Only 1–2% of the genomes encodes proteins, so there would seem to be an additional common set (about 3%) of functional elements in mammalian non-coding DNA. These common sequences may constitute, for example, regulatory elements, structural elements or RNA genes. Notably, such regions are found mostly within the 0.8 Gbp of ancestral sequence common to human, mouse and dog.

With the dog genome sequence available, it will be exciting to follow the forthcoming search for associations between certain phenotypes in different breeds and the genes responsible for them (Box 1). It will now be possible, using various genomic approaches, to map breed-specific traits related to morphology, physiology and behaviour¹⁵, which will provide insight into key features of mammalian biology and disease. ■

Hans Ellegren is in the Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden.
e-mail: Hans.Ellegren@ebc.uu.se

2. Vilà, C. *et al.* *Science* **276**, 1687–1689 (1997).

3. Savolainen, P. *et al.* *Science* **298**, 1610–1613 (2002).

4. Leonard, J. A. *et al.* *Science* **298**, 1613–1616 (2002).

5. Sutter, N. B. & Ostrander, E. *Nature Rev. Genet.* **5**, 900–910 (2004).

6. Lindblad-Toh, K. *et al.* *Nature* **438**, 803–819 (2005).

7. International Human Genome Sequencing Consortium *Nature* **431**, 931–945 (2004).

8. Kirkness, E. F. *et al.* *Science* **301**, 1898–1903 (2003).

9. Bruford, M. *et al.* *Nature Rev. Genet.* **4**, 900–910 (2003).

10. Vilà, C. *et al.* *Trends Genet.* **21**, 214–218 (2005).

11. Götherström, A. *et al.* *Proc. R. Soc. Lond. B* **272**, 2337–2344 (2005).

12. Sutter, N. B. *et al.* *Genome Res.* **14**, 2388–2396 (2004).

13. Parker, H. G. *et al.* *Science* **304**, 1160–1164 (2004).

14. www.ncbi.nih.gov/Genomes

15. Pollinger, J. P. *et al.* *Genome Res.* (in the press).

16. Lin, L. *et al.* *Cell* **98**, 365–376 (1999).

17. Lohi, H. *et al.* *Science* **307**, 81 (2005).

WATER

Ins and outs of ice nucleation

Srikanth Sastry

Laboratory experiments point to a mechanism by which ice forms from supercooled water with surprising alacrity. Such a mechanism may help to explain ice formation in the atmosphere under certain conditions.

In Kurt Vonnegut's novel *Cat's Cradle*, Earth's waters freeze over on contact with ice-9, a fictional form of ice that is more stable than water. Vonnegut depicts an imaginary and extreme scenario of how the thermodynamics of water might dictate the fate of the planet. But the transformation of water to ice is no less fascinating in reality. Water can remain a liquid even under conditions in which a more stable phase exists. This occurs at temperatures below 0 °C when ice-I (the normal variety) is the stable phase. In these circumstances, water is termed metastable or supercooled. It can be prompted to turn into ice by seeding it with a speck of the stable ice (as in the story), or with a small particle of, for instance, dust or ash.

At Earth's surface, water exists in solid, liquid and gaseous phases at or close to ambient conditions. Transformations between these phases can therefore occur readily (through small changes in temperature, for example), and can have a great influence on the dynamics of Earth's atmosphere. This is the context in which two new publications by Raymond A. Shaw and colleagues^{1,2} are set. The authors describe lab studies of a particular pathway, called contact nucleation, by which

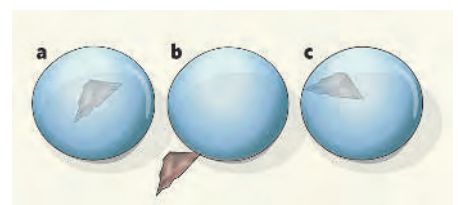


Figure 1 | Three ways in which an ice nucleus may cause crystallization of a water drop.
a, A nucleus immersed in the bulk drop.
b, Contact from a nucleus outside the drop.
c, Contact from within the drop ('contact nucleation inside-out'). Crystallization occurs at higher temperatures in the two surface-contact situations^{1,2}.

supercooled water can be transformed into ice. They also discuss how their results might help to understand aspects of ice formation in Earth's atmosphere — which in turn affects patterns of rainfall and snowfall, and the influence of clouds on the amount of solar radiation reaching Earth's surface.

When cooled below 0 °C at ambient pressure, water will eventually become ice, a solid with a regular molecular structure and strong attractive interactions between molecules that

1. Coppinger, R. & Coppinger, L. *Dogs: A Startling New Understanding of Canine Origin, Behaviour & Evolution* (Scribner, New York, 2001).



Figure 2 | Ice nucleation in wave clouds. Wave clouds — shown here — form when air is lifted up over a mountain, and water vapour in the upper reaches of the air current condenses to form water droplets. As the air current descends, the condensed water evaporates. Such an air current can bounce up and down, and with condensation at the crests of such undulations a wave-like cloud pattern can emerge. Durant and Shaw² argue that ‘contact nucleation inside-out’ may explain the extent of ice formation occurring in the downwind region of such currents.

becomes the thermodynamically preferred state. But the transformation first requires a germ of the crystal, a crystallite, to form. Such crystallites are created spontaneously as a result of the incessant jiggling around of molecules in the liquid, caused by thermal fluctuations. However, most of the crystallites dissolve back into the liquid. To grow, crystallites must overcome the barrier posed by surface tension — the unfavourable interactions between molecules in the crystalline arrangement and those in the surrounding liquid. A large enough crystallite that can overcome the barrier is termed the critical nucleus; the process by which it forms is called nucleation. Once formed, the critical nucleus will grow irreversibly and ice will form.

Water can be supercooled because crystal nucleation and growth may take a long time. But the processes become faster as the temperature falls, and they can be accelerated further by solid particles that act as a substrate for crystal nucleation. This is known as heterogeneous nucleation. It is of particular importance in the atmosphere, where water exists as small droplets that are suspended alongside particles that may act as heterogeneous ice nuclei. These nuclei may either be immersed in the bulk of the water droplet, or come into contact with the droplet surface.

Shaw and co-workers have examined the phenomenon of contact nucleation in lab experiments. As discussed in the first paper¹, they set out to compare the efficiency of an ice nucleus in causing crystallization when it is immersed in the droplet and when it is in contact with the droplet's surface. They used an experimental set-up in which the same drop of water, with the same ice nucleus immersed or in contact, is repeatedly cooled and heated hundreds of times. During each cooling run, they recorded the temperature at which the droplet froze. The result is an estimate of the most likely temperature at which droplets

freeze, which in turn is a measure of the efficiency of the particular nucleation mechanism.

They find that the freezing temperature when the ice nucleus is in contact with the droplet is about 5 °C higher than when it is immersed in the droplet, showing that contact nucleation is a more effective mechanism than immersion nucleation. This conclusion is consistent with previous data. But the new results show that the efficacy of contact nucleation is not caused by transient effects related to an ice nucleus coming into contact with a water droplet, such as mechanical disturbance due to collision, or to the dissolution of part of the ice nucleus. Instead, it has simply to do with the fact that the nucleus is in contact with the droplet surface. This is a useful distinction, which may also be related to a proposal³ that homogeneous nucleation (nucleation without an ice nucleus) is most effectively initiated at the droplet surface.

In the second paper², Durant and Shaw

describe a variation of the experiment in which the ice nucleus is initially immersed in the droplet, but in conditions under which the droplet slowly evaporates as it is cooled and heated. The nucleus then eventually comes into contact with the surface of the droplet, but this time from inside the droplet (Fig. 1). In this situation, too, the authors observe a rise in the freezing temperature.

This ‘contact nucleation inside-out’ is evidently an efficient nucleation mechanism. But is it of special significance? Yes, claim the authors. They suggest that it may account for the high rates of ice nucleation in wave clouds (Fig. 2), when the cloud droplets are evaporating at temperatures that are too high for the rates to be explained by homogeneous nucleation. Re-examination of existing observational data may provide support for this idea.

Shaw and colleagues' results^{1,2} will need to be validated, but they are appealing in their clarity and possible relevance to ice nucleation in the atmosphere. Extension of the experiments, perhaps with a clever choice of ice nuclei and varying droplet sizes, may also provide further evidence about homogeneous surface nucleation⁴. Another route forwards is through computer simulations, which have provided insight into heterogeneous nucleation⁵ and homogeneous crystal nucleation in water⁶. Such simulations could be extended to provide a fresh angle on homogeneous and heterogeneous nucleation at surfaces. ■

Srikanth Sastry is at the Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore 560064, India.
e-mail: sastry@jncasr.ac.in

1. Shaw, R. A., Durant, A. J. & Mi, Y. J. *Phys. Chem. B* **109**, 9865–9868 (2005).
2. Durant, A. J. & Shaw, R. A. *Geophys. Res. Lett.* **32**, doi:10.1029/2005GL024175 (2005).
3. Tabazadeh, A., Djikaev, Y. S. & Reiss, H. *Proc. Natl Acad. Sci. USA* **99**, 15873–15878 (2002).
4. Kay, J. E. et al. *Atmos. Chem. Phys. Discuss.* **3**, 3361–3372 (2003).
5. Cacciuto, A. et al. *Nature* **428**, 404–406 (2004).
6. Matsumoto, M. et al. *Nature* **416**, 409–413 (2002).

CELL BIOLOGY

Relays at the membrane

Roel Nusse

The Wnt signalling pathway is a major route by which the cell conveys information from its exterior to the nucleus. A gap in the sequence of signalling proteins has now been filled.

The process of signal transduction allows a cell to receive messages from its environment and transfer this signal from the membrane through the cytoplasm and into the nucleus. Here the signal alters the expression of the various genes that contribute to the cell's response. Regardless of the signal's nature, the general logic of the transduction pathways that

are triggered by protein ligands is roughly the same. The signalling protein binds to a receptor on the cell's surface, which consequently undergoes a conformational change. Commonly, the receptor is then tagged with a phosphate group by an associated protein kinase enzyme. The phosphorylation allows the receptor to recruit cytoplasmic signalling

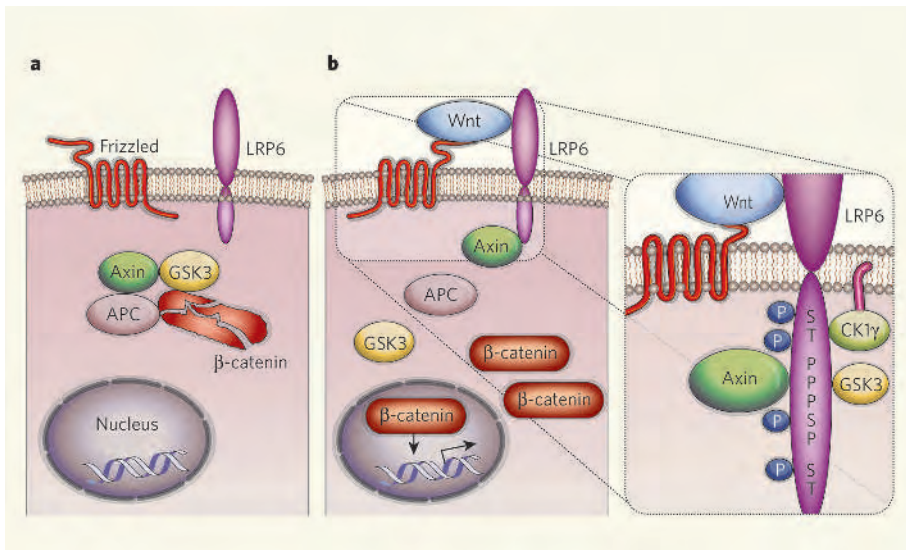


Figure 1 | Crucial kinases. **a**, In cells not activated by Wnt, a complex between β -catenin, Axin, APC and GSK3 causes phosphorylation of β -catenin and its consequent destruction. The Wnt receptors LRP6 and Frizzled are unoccupied. **b**, Without Axin, β -catenin is stabilized and it enters the nucleus to control gene expression. Inset, binding of Wnt to cells results in phosphorylation (P) of LRP6 residues in its cytoplasmic tail. Zeng *et al.*¹ and Davidson *et al.*² show that this is catalysed by the GSK3 and CK1 γ protein kinases. CK1 γ is attached to the membrane by a lipid anchor domain. Several other sites on LRP6 that become phosphorylated are not shown here. The phosphorylated LRP6 recruits Axin, removing it from the β -catenin destruction complex and stabilizing β -catenin.

components that initiate a cascade of events resulting in changes in gene expression. Two papers in this issue^{1,2} show that signal transduction initiated by the protein Wnt — a major regulator of developmental processes — follows a similar strategy but with some interesting new twists.

Compared with other signalling cascades, the Wnt pathway³ is exceptional in its complexity, with numerous components and intricacies that go beyond this short overview. One of the key players in the pathway is β -catenin, a protein that resides in the cytoplasm and, once activated, is responsible for relaying the signal into the nucleus. When cells are not exposed to Wnt, β -catenin is destined to be destroyed (Fig. 1a). This process is triggered by phosphorylation of β -catenin, catalysed by the protein kinase GSK3 and assisted by the β -catenin-binding partners Axin and APC. The Wnt signal activates two membrane receptors, the Frizzled and LRP6 molecules, which form a complex and trigger signalling to the cytoplasm that halts the breakdown of β -catenin⁴. But how are the events at the receptor level coupled to β -catenin, and how does β -catenin escape degradation?

The discovery, several years ago, that the Axin protein can bind to the cytoplasmic tail of the LRP6 receptor⁵, provided a mechanism by which Axin is seized from β -catenin⁶. This changes the fate of β -catenin: instead of being destroyed, it accumulates and enters the nucleus, where it executes a programme of Wnt-induced gene expression (Fig. 1b). Crucially, the binding of Axin to the LRP6 tail is promoted by phosphorylation of LRP6 (ref. 7), suggesting that protein kinases must be recruited to the receptor after activation by Wnt⁸.

The identity of these kinases is the subject of the papers by Zeng *et al.* (page 873 of this issue)¹ and Davidson *et al.* (page 867)². To appreciate their function, some detail is necessary. Phosphorylation of LRP6 occurs on several clusters of serines and threonines, with a central proline-rich motif (PPPSP) as a hallmark (Fig. 1b; inset). As in many other cases of cluster phosphorylation, there is a priming phosphorylation event after which the remaining residues become modified as well. Proline-rich environments are conducive to phosphorylation by GSK3, so Zeng *et al.*¹ tested the PPPSP motif on LRP6 for activity. They found that the serine in the motif is indeed modified by GSK3, leading to activation of signalling. Strikingly, GSK3 is now known to phosphorylate a number of Wnt signalling components, including β -catenin, Axin and APC. GSK3 used to be thought of as a negative component in Wnt signalling: when it was deleted genetically, Wnt-response genes were activated because β -catenin was no longer phosphorylated or degraded. But we now know that it acts positively on LRP6, activating Wnt signalling — an effect missed in the genetic experiments because of its negative involvement further down the pathway.

Residues next to the PPPSP motif also get phosphorylated. What is the enzyme? Based on an expression screen, Davidson *et al.*² identify this kinase as a member of the CK1 family, CK1 γ . Beyond biochemical experiments showing that CK1 can phosphorylate the LRP6 tail, Davidson *et al.* demonstrate that the gene encoding CK1 γ is required for Wnt signalling to occur, and that overexpression of this gene is sufficient to activate the pathway.

As with GSK3, the diminutive name CK1 does not do justice to the numerous functions of the CK1 family in cell physiology. Within this family, CK1 γ is an outlying relative, and, interestingly, it has a membrane anchor in the form of a fatty-acid attachment site. Eliminating the fatty-acid anchor domain from CK1 γ results in loss of Wnt signalling, implying that this kinase needs to be associated with the membrane to act.

So now there are two LRP6 kinases, raising the question of how these enzymes become activated by the Wnt signal. Here the two papers differ in their conclusions. Zeng *et al.* suggest that the GSK3-dependent phosphorylation of the PPPSP motif is induced by Wnt. By contrast, Davidson *et al.* conclude that PPPSP is usually phosphorylated in cells, in the absence of Wnt; that is, it is constitutively phosphorylated. They propose that it is the subsequent modification of neighbouring residues, catalysed by CK1 γ , that is dependent on the Wnt signal. This discrepancy needs to be resolved, but if phosphorylation of LRP6 by GSK3 is indeed signal dependent, it would be an exception to the general rule that GSK3 activity is constitutive in cells. GSK3 is involved in many signalling pathways, but it acts on all its multitude of targets without being triggered by a signal from outside. By contrast, CK1 γ activity is clearly stimulated by Wnt signalling, as adding Wnt protein to cells leads to modification within a few minutes.

There are many questions remaining. For example, it is not known how CK1 γ activity is regulated or whether the enzyme becomes physically associated with LRP6. Because of its unique function, this enzyme provides an attractive novel target for Wnt-specific inhibitors.

When we now compare Wnt signalling events at the receptor level to other signalling pathways, there are many parallels but one difference. In the Wnt signalling pathway, ligand binding triggers the formation of a receptor complex, and protein kinases modify the receptor tails, leading to recruitment of cytoplasmic factors. In other signalling pathways, however, receptor-induced protein phosphorylation amplifies the signal, and the receptor-associated kinase acts as a catalyst for the modification of many substrate molecules. In this regard, Wnt signalling is peculiar: Wnt-induced LRP6 phosphorylation acts by titrating away a negative regulator of signalling, Axin. This implies a stoichiometric rather than a catalytic mechanism of signal transduction. On the other hand, Axin is present in very low concentrations in cells, much lower than the other components in the β -catenin destruction complex⁹. So, is it possible that Axin actually plays a dynamic role, shuttling between the receptor and the destruction complex and acting as an amplifier of Wnt signalling rather than as a simple scaffold?

Roel Nusse is in the Howard Hughes Medical Institute, Department of Developmental Biology,

Beckman Center, Stanford University, Stanford, California 94305-5329, USA.
e-mail: rnusse@stanford.edu

1. Zeng, X. *et al.* *Nature* **438**, 873–877 (2005).
2. Davidson, G. *et al.* *Nature* **438**, 867–872 (2005).
3. Logan, C. & Nusse, R. *Annu. Rev. Cell Dev. Biol.* **20**, 781–810 (2004).

4. Cong, F., Schweizer, L. & Varmus, H. *Development* **131**, 5103–5115 (2004).
5. Mao, J. *et al.* *Mol. Cell* **7**, 801–809 (2001).
6. Tolwinski, N. S. *et al.* *Dev. Cell* **4**, 407–418 (2003).
7. Tamai, K. *et al.* *Mol. Cell* **13**, 149–156 (2004).
8. He, X., Semenov, M., Tamai, K. & Zeng, X. *Development* **131**, 1663–1677 (2004).
9. Lee, E., Salic, A., Kruger, R., Heinrich, R. & Kirschner, M. W. *PLoS Biol.* **1**, E10 (2003).

QUANTUM INFORMATION

Remember that photon

Philippe Grangier

Storing single photons in atomic memories, and releasing them at a later time, is a required step on the way to quantum repeaters and long-distance quantum cryptography networks. This step has now been taken.

The basic unit of quantum information, the quantum bit or qubit, can be encoded in various physical quantities, such as the polarization states of photons, or the spin states of atomic nuclei. To make qubits practically useful, random coupling of them with the external world — an effect known as decoherence — must at all costs be avoided or corrected. This makes photons (the quanta of light) particularly suitable for qubit transmission, as they can travel over very long distances with very little decoherence. For qubit storage, encoders such as atoms come into their own: they can be kept in 'traps' for long periods, again avoiding deleterious decoherence effects from outside.

In experiments detailed in two papers in this issue, Chaneilère *et al.* (page 833)¹ and Eisaman *et al.* (page 837)² contrive to combine the two crucial aspects of transport and storage: they generate a single photon on demand, catch it and store it in a remote atomic memory, and release it some time later. The advance is potentially highly significant for the field of quantum cryptography, also known as quantum key distribution (QKD). This emerging technology promises absolutely secure transmission of the key codes that are essential to decipher any encrypted message (Box 1).

Previous advances in quantum key distribution have owed much to the fact that photons that are used to encode the keys are very good qubit carriers: apart from maintaining a robust quantum state throughout transmission, they can be detected efficiently and with low levels of noise. But light signals cannot — whether viewed classically or quantum-mechanically — propagate over infinite distances in optical fibres. They are in fact dampened exponentially with distance: by a factor of two over 15 kilometres, and by a factor of a hundred over 100 kilometres. In classical optical telecommunications, this problem is solved by using simple, readily available devices known as repeaters, which can amplify and reshape the transmitted signal. But a good classical

repeater is no use in the quantum regime: it is much too noisy, and creates so many errors that any quantum key being transmitted would not survive. To put the problem in more quantum-mechanical terms, a classical repeater breaks down quantum entanglement. This delicate phenomenon is associated with very strong, non-classical correlations between the states of two widely separated qubits, and is a crucial element of all quantum communication schemes: in effect, it allows any useful qubit to be 'teleported' directly to its destination, avoiding transmission losses³.

So quantum communication must reinvent the repeater concept, using quantum hardware that preserves coherence. This is feasible in principle⁴: a quantum repeater would be nothing more than a small quantum processor. The exact number of qubits that would have to be stored and processed in such a repeater to ensure high-fidelity quantum communication over thousands of kilometres is an open issue. But it is likely to be in the

range of tens or hundreds — much lower than the number required for a fully fledged quantum computer. The proposal in 2001 of the so-called DLCZ quantum information protocol⁵, in which an ensemble of many atoms stores just one qubit, was a significant step towards a functioning quantum repeater. This protocol uses a process known as spontaneous Raman scattering, in which an incident photon is scattered inelastically (that is, with a change in its frequency) between two atomic ground states.

Chaneilère *et al.*¹ and Eisaman *et al.*² exploit the DLCZ protocol to set up a controllable single-photon source for further experimentation. After initially preparing all the atoms of an ensemble in one ground state, a weak laser pulse (which nevertheless contains many photons) is used to induce a Raman transition of just one atom within the ensemble. As a consequence, a single spontaneous Raman photon is scattered, and its detection heralds the creation of a collective, delocalized, single-atom excitation of the ensemble. This excitation can be stored for as long as all the atomic levels in the sample maintain a constant phase relationship (a period known as the coherence time of the ensemble). This excitation can be converted back into a single-photon light field of controllable direction, intensity and frequency using another pump pulse (for a review of recent experimental work in this area, see ref. 6).

Once a single photon has been generated, the second stage is to catch it, and then release it again, in a second, remote atomic ensemble. The trick here is to use a second atomic ensemble that is opaque to the photon — absorbing rather than transmitting it — and that can only be made transparent by using an extra laser beam. This transparency arises through a neat and extensively studied interference phenomenon, electromagnetically induced transparency (EIT). If the EIT laser

Box 1 | Key codes: classical versus quantum cryptography

The purpose of quantum key distribution is to share a secret key among legitimate users that allows them and only them to decode messages. Some sort of key that allows a message to be deciphered is essential to all forms of encryption. Common, classical schemes used in electronic commerce can set up a key by relying on computationally difficult problems, such as the splitting of a very large number into two prime-number factors, that are in fact — given unlimited patience and computational power — breakable.

The only totally secure classical encryption system is the 'one-time pad', which uses a key that is as long as the message itself and that may be used only once. This solution leads to what is known as the key distribution problem: as the key must be transmitted between sender and recipient, it is itself susceptible to interception by an eavesdropper. In the classical world, someone

can listen in on such a signal passively without changing the bits that make it up at all, so neither sender nor recipient need ever know that their communication has been intercepted.

Not so in the world of quantum communication. Qubits do not possess definite values such as the 0 or 1 of classical bits; rather, they represent a so-called coherent superposition of physical states such as the polarizations of a photon. A fundamental feature of quantum mechanics is that the mere act of observing such a superposition will cause it to 'collapse' into a definite state. This means any attempt by an eavesdropper to intercept a key made of qubits can be easily spotted by sender and recipient. Given this knowledge, and as long as the errors created by the eavesdropper (or any other perturbation) are not too large, it should be possible to build up an errorless and perfectly secure key.

P.G.



50 YEARS AGO

Over the past twenty years I have been interested in the possibility of using X-ray crystallographic methods to find the arrangement of the atoms in protein molecules and particularly in insulin. One of many possible approaches to solving this problem seems to be the crystallographic study of naturally occurring peptides such as the gramicidins and tyrocidine... These all have molecules much smaller in size than even the smallest protein molecules; some indeed are smaller than vitamin B₁₂, of which we have already found it possible to obtain the kind of information we require... We already have evidence that there may be a connexion between the way the peptide chain is folded in gramicidin S and the way it is folded in part of the molecule of insulin. But even if later we find that the connexion in chain configuration is less close than we at present suppose, we think that the atomic arrangement in these peptide molecules is itself of great interest and some importance.

Dorothy Crowfoot Hodgkin
From *Nature* 10 December 1955.

100 YEARS AGO

The death-knell of the atom¹

Old Time is a-flying; the atoms are dying;
Come list to their parting oration:—
“We’ll soon disappear to a heavenly sphere
On account of our disintegration.

“Our action’s spontaneous in atoms uranious
Or radious, actinious or thorious:
But for others, the gleam of a heaven-sent beam
Must encourage their efforts laborious.

“For many a day we’ve been slipping away
While the savants still dozed in their slumbers;
Till at last came a man with gold-leaf and tin can
And detected our infinite numbers.”

¹Sung at the Chemical Laboratory dinner at University College, November 17.

From *Nature* 7 December 1905.

beam is turned off, the medium becomes opaque once again, and any photon inside it is trapped, converting into another atomic excitation (known as a dark-state polariton). The photon can be regenerated at any time within the coherence time of the ensemble, simply by turning the EIT laser beam on again.

Throughout this sequence of events, it is clearly essential to check that the photon maintains its quantum, particle-like properties. One way to do this is to put a beam-splitter in the photon’s way, and verify that photon counts on the two paths after the beam-splitter are anticorrelated. Correlated counts would indicate that the incoming beam splits in two, a clear sign of classical, wave-like behaviour. The degree of photon splitting can be conveniently characterized⁷ by a parameter α , with an ideal single photon (exhibiting a purely quantum behaviour) having $\alpha = 0$, and a classical source having $\alpha > 1$. A value of α between 0 and 1 thus corresponds to a light beam showing a mixture of quantum and classical behaviours, or in other words, to an imperfect single photon. Chaneilère and colleagues¹ obtain a value for α of 0.36 after a storage time of 500 nanoseconds, whereas Eisaman and colleagues² find a value of 0.51 under EIT conditions, but without storage (they also observe storage, but without evidence that α is less than 1).

Obviously, the ‘quantum memories’ (the ability to ‘regenerate’ a photon stored in an ensemble after a delay) described in these articles^{1,2} are not the end of the story. First, what has to be stored and released is not a photon, but a qubit — the quantum information encoded on a photon. In the context of the

DLCZ proposal, how to store and release a qubit is known in principle, and preliminary results have been obtained⁸. Another crucial issue is that it should be possible to create some entanglement between distant atomic ensembles⁹, as discussed also by Chou *et al.* in this issue (page 828)¹⁰. The next key step will be to gradually increase the entanglement between the two remote memories — the process of ‘entanglement distillation’, which would be the fundamental duty of a quantum repeater^{3,4}. This is a long-term goal, as many features have to be improved: counting rates (presently much too low), storage times (presently much too short), and the fidelity of the successive transfer processes in the ensembles. Although this looks more like mountain climbing than highway driving, new ways upwards keep on opening, as the present research^{1,2} shows. The summit may seem far off, but it is not out of reach. ■

Philippe Grangier is at the Laboratoire Charles Fabry de l’Institut d’Optique, 91403 Orsay Cedex, France.

e-mail: philippe.grangier@iota.u-psud.fr

1. Chaneilère, T. *et al.* *Nature* **438**, 833–836 (2005).
2. Eisaman, M. D. *et al.* *Nature* **438**, 837–841 (2005).
3. Bennett, C. H. *et al.* *Phys. Rev. Lett.* **70**, 1895 (1993).
4. Briegel, H.-J., Dür, W., Cirac, J. I. & Zoller, P. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
5. Duan, L.-M., Lukin, M., Cirac, J. I. & Zoller, P. *Nature* **414**, 413–418 (2001).
6. Felinto, D. *et al.* *Phys. Rev. A* **72**, 053809 (2005).
7. Grangier, P., Roger, G. & Aspect, A. *Europhys. Lett.* **1**, 173–179 (1986).
8. Matsukevich, D. N. *et al.* *Phys. Rev. Lett.* **95**, 040405 (2005).
9. Matsukevich, D. N. *et al.* preprint available at www.arxiv.org/quant-ph/0511012 (2005).
10. Chou, C. W. *et al.* *Nature* **438**, 828–832 (2005).

CANCER BIOLOGY

Emissaries set up new sites

Patricia S. Steeg

The capacity of tumours to spread to other organs is one of their most dangerous attributes. A study of how cancer cells settle in new places shows that they send out envoys to prepare the ground for them.

During the process of metastasis, tumour cells move from the primary tumour to colonize another organ. But why do these mobile cells put down roots only in particular organs, or only at specific sites within an organ? The lungs and liver, for example, seem particularly popular secondary targets for tumour cells. Some studies imply that this ‘preference’ might occur because, as they branch out within those organs, the blood vessels become very narrow, and the blood-borne tumour cells are trapped when they enter the fine capillary beds¹. Other work has identified proteins that are specific to the cells lining the capillaries of certain tissues as possibly promoting metastasis formation².

A report from Kaplan *et al.* (page 820 of this issue)³ provides another explanation. The authors show that tumour cells can mobilize normal bone-marrow cells, causing them to migrate to particular regions and change the local environment so as to attract and support a developing metastasis.

Metastasis is a sequential process, contingent on tumour cells breaking off from the primary tumour, travelling through the bloodstream, and stopping at a distant site. At the new site, the cells establish a blood supply and can grow to form a life-threatening mass. Both stimulatory and inhibitory molecular pathways within the tumour cell regulate this

behaviour^{4,5}, and interactions between the tumour cell and host cells in the distant site are also significant. One of the best examples is the vicious cycle in bone metastasis, where tumour cells secrete parathyroid-hormone-related protein (PTHrP), which in turn activates normal bone cells to break down the bone matrix. This degradation releases embedded factors⁶, such as transforming growth factor- β (TGF- β), that stimulate the tumour cells to proliferate and secrete more PTHrP. But are other cells involved in the metastatic process? And what are the earliest events at the metastatic site?

Kaplan *et al.*³ set up an ingenious experiment to track the movements of various cell populations as tumour cells metastasized in the lungs of live mice (Fig. 1). The mice were irradiated to kill off all their bone-marrow cells, which were then replaced by bone-marrow cells tagged with green fluorescent protein; this made the cells easy to find under a microscope. Once the new bone-marrow cells were established, the mice were injected in the skin with lung carcinoma or melanoma cells, each marked with red fluorescent protein.

The tumour cells were expected to form a primary tumour in the skin, and then to metastasize to the lungs. But the green bone-marrow-derived cells appeared in the lungs on days 12–14 after injection of the red cells — well before any of the tumour cells had arrived in the lung. The red tumour cells turned up only on day 18 post-injection, and by day 23 micrometastases had formed, with more than 95% of the tumour cells being found in exactly the same sites as the bone-marrow-derived cells.

In a variation on this experimental theme, the authors injected the mice with the medium in which the melanoma cells had been cultured, rather than with the cells themselves. This also caused the bone-marrow cells to move to the animals' lungs, implying that the melanoma cells had secreted some factor into the surrounding solution that mobilized the bone-marrow cells. The mice were subsequently given red-tagged melanoma cells intravenously, and four days later 93% of these cells were found together with the bone-marrow cells in the lungs.

In the jargon of cancer biologists, tumours exist in a 'niche'; this is analogous to an organism living in an ecological niche, in that the cancer is adapted to — and taking advantage of — its local physiological environment. Kaplan *et al.* explain their results by proposing that the tumour cells act to set up a pre-metastatic niche ready for their arrival in the lungs, sending bone-marrow cells in first to create a suitable environment for the tumour cells to settle in.

The bone-marrow-derived cells used in the experiment were a mixture of cell types and developmental stages. So could all the cells produce pre-metastatic niches, or only

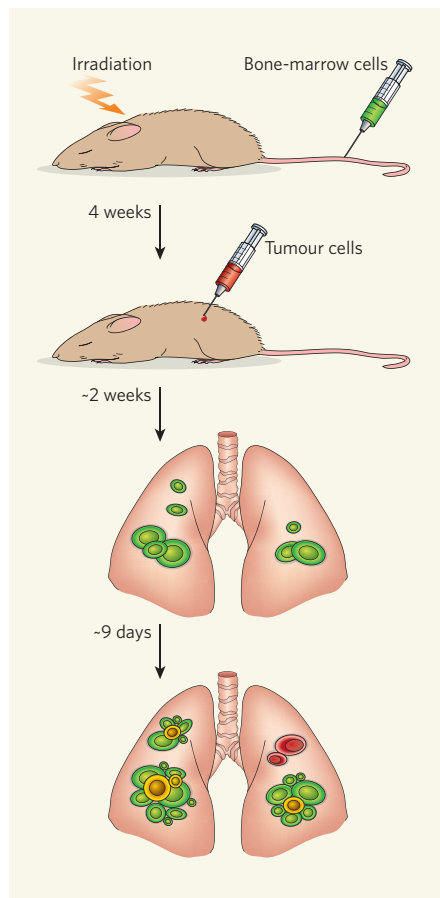


Figure 1 | The experiments of Kaplan and colleagues³. They find that cells derived from the bone marrow (green) precede tumour cells (red) to the lung, the site of metastasis. The bone-marrow cells create a proposed 'pre-metastatic niche', and the tumour cells join them to form a metastasis (yellow; in fluorescence microscopy, green overlaid with red produces a yellow colour).

a specific subset of them? The bone-marrow cells expressed a number of proteins characteristic of haematopoietic progenitor cells — immature cells that can divide and produce mature blood and bone-marrow cells — in particular, vascular endothelial growth factor receptor 1 (VEGFR1). Treatment of mice with an antibody to VEGFR1 prevented the pre-metastatic niches from forming. Moreover, if the experiment was carried out with bone-marrow cells that did not express VEGFR1, neither pre-metastatic niches nor metastases formed. The authors next examined a small set of primary human tumours and lymph-node metastases from several cancer types, and found cells with VEGFR1 present in these, too. They propose that these might also be pre-metastatic niche cells.

So what is going on? Kaplan *et al.* propose a series of pathways, although additional insights will undoubtedly follow. Molecular factors secreted by the tumour cells seemed to stimulate normal fibroblast cells in the future metastatic site to produce fibronectin, a protein that binds cells to the extracellular matrix

that holds them together to form tissues. If so, this might create a 'docking site' for the arriving bone-marrow cells. Consistent with this idea, the bone-marrow cell population expressed $\alpha_4\beta_1$ and $\alpha_4\beta_7$ integrins, which are proteins that bind to fibronectin to mediate cell–cell attachments. Protease production by the bone-marrow cells may liberate growth factors (including VEGF) that are embedded in the extracellular matrix to support the developing niche. Using *in vitro* experiments, the authors showed that the VEGFR1-positive bone-marrow cells promoted the attachment and motility of tumour cells. They propose that similar interactions may operate *in vivo* to form the pre-metastatic niche and eventually the micrometastasis.

The role of bone-marrow-derived cells in tumour formation and the growth of new vasculature in tumours has been controversial⁷. How these subpopulations are linked to the subpopulation that mediates pre-metastatic niches is not known. But it is intriguing that the list of signals that stimulate bone-marrow progenitor cells mostly overlaps that for metastatic cancer cells, including integrins and other attachment proteins. TGF- β , proteases and VEGF. Although VEGFR1 seems to be essential for pre-metastatic niche formation by the bone-marrow-derived cells, several other activators of this receptor, including VEGF and placental growth factor, may be involved.

Once these findings have been confirmed and extended to elucidate the factors concerned, inhibitors of the pathway would be of great interest for potentially blocking metastasis. But if such preventative therapy were to be tested in the clinic, it should not be directed at patients with metastatic disease, which is where most clinical testing begins. In such patients, the pre-metastatic niches and metastases have already formed, and it is doubtful that a compound would reverse the process. Rather, an anti-VEGFR or similar compound would be most germane to patients in the 'adjuvant setting', for instance those at high risk of metastatic disease because tumour cells have already invaded the regional lymph nodes. It is in these patients that metastases may be forming, and might be interrupted. The development of molecular metastasis blockers may therefore force a redesign of the clinical-trials testing system to accommodate the molecular biology of metastatic disease. ■

Patricia S. Steeg is in the Women's Cancers Section, Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, USA. e-mail: steegp@mail.nih.gov

1. Weiss, L. *et al.* *J. Pathol.* **150**, 195–203 (1986).
2. Pasqualini, R. & Ruoslahti, E. *Nature* **380**, 364–366 (1996).
3. Kaplan, R. N. *et al.* *Nature* **438**, 820–827 (2005).
4. Steeg, P. *Nature Rev. Cancer* **3**, 55–63 (2003).
5. Sawyer, T. K. *Expert Opin. Invest. Drugs* **13**, 1–19 (2004).
6. Mundy, G. *Nature Rev. Cancer* **2**, 584–593 (2002).
7. Garmy-Susini, B. & Varner, J. A. *Br. J. Cancer* **93**, 855–858 (2005).

OBITUARY

Alastair Cameron (1925–2005)

Astrophysicist and planetary scientist.

Almost all our theories of the origin of the stars, the planets and the chemical elements have been developed in the past half-century. These were the diverse arenas in which Alastair G. W. Cameron worked, and where his ideas — he was a well-spring of ideas — inspired and guided two generations of colleagues.

Cameron, who died on 3 October in Tucson, Arizona, was born in Winnipeg, Canada. He received his PhD in nuclear physics from the University of Saskatchewan in 1952, and took a faculty post at Iowa State College in Ames. While there, he read of the discovery of the highly unstable element technetium in the atmospheres of a certain class of red-giant star. He was inspired by this to wonder about the source of the flood of neutrons that would be required to maintain a spectroscopically observable concentration of the element. Ignorant of astrophysics, Cameron acquired a shelf of textbooks and set about educating himself. Without regret, he thus turned his back on experimental physics, embarking on a lifetime career as a theorist. While at Iowa, Cameron also met Elizabeth MacMillan at a science-fiction convention; she would become his wife.

In 1954, Cameron joined the Canadian atomic-energy project at Chalk River, Ontario. There he worked on calculating the reaction rates that control nucleosynthesis inside stars. (Nucleosynthesis is the creation of chemical elements through nuclear processes such as fission, fusion and neutron capture.) After a brief stay at the California Institute of Technology in Pasadena, he moved in 1961 to the NASA Goddard Institute for Space Studies in New York City's Upper West Side. Given Cameron's increasing interest in the creation of the Solar System, the fact that the institute's home was called the Interchurch Building was a source of mild amusement for his colleagues.

Cameron's approach to understanding the formation of chemical elements was guided by their present-day abundances in the Solar System, which had been reviewed in 1956 by Hans E. Suess and Harold C. Urey. In working to update and improve their table, Cameron came to appreciate the importance of data from meteorites. Seeing the need to improve communication between the meteoritic and astrophysical communities, he instituted a series of interdisciplinary meetings in New England under the aegis of the Gordon Research Conferences, first

during the 1960s and again in the 1990s. In the 1960s, Cameron also wrote his first semi-quantitative papers modelling the solar nebula, the disk of matter from which the meteorites and planets of the Solar System formed. These papers came to be taken very seriously by the meteoritics community — perhaps more seriously than Cameron, who saw them rather as steps towards the right answer, intended them to be taken.

In 1966, Cameron (now a US citizen) moved to the Belfer Graduate School of Science at Yeshiva University in Manhattan, and in 1973 became a professor at Harvard University and associate director of the Harvard-Smithsonian Center for Astrophysics. He remained at Harvard until his 'retirement' in 1999 — when he shifted his activities, undiminished, to the Lunar and Planetary Laboratory at the University of Arizona.

Any account of Cameron's career would be incomplete without noting his love affair with digital computers. The complex, interlocking nonlinear equations that describe the reaction networks in stellar nucleosynthesis and the behaviour of protostellar disks require numerical integration, which in turn needs computing power. Cameron had been an eager consumer of that commodity ever since his first encounter with a card-fed IBM 650 in his Chalk River days. As soon as he could, he acquired his own minicomputers, preferring these to the shared mainframes at his various institutions. Cost-effectiveness was his justification; in truth, it just pleased him to be the master of his own computing. And no one was more knowledgeable on the subject: for about a decade, computers dominated his social conversation. Visitors to his office frequently complained that the whine of massed cooling fans made conversation difficult.

But Cameron was a man of uncommonly broad interests, and his activities and enthusiasm routinely overflowed the boundaries that confine most workers. From discussions of the solar nebula it was a short step to questions of planetary formation, and so to the origin of the Moon (which had not been explained by the results of the Apollo Moon missions of the 1960s and '70s). Cameron and others suspected that a gigantic collision between sub-planets during Earth's accretion must have blasted off debris, creating an orbiting circumterrestrial disk from which the Moon formed.

The proposition was difficult to test quantitatively: the trajectories of a huge



number of debris fragments needed to be followed, in a complex gravitational field and under the effects of an expanding vapour cloud. The problem posed a fine challenge to Cameron's computational prowess, and he rose to it royally, making clear in a series of co-authored papers dating from 1976 that the Moon could have been created in a collision. In the light of the severe problems attached to other models, it became generally accepted that this was indeed the case.

This was arguably the closest Cameron came to reaching closure on a major problem. His many papers on nuclear astrophysics (his first and continuing love), star formation, and protostellar-disk theory contributed profoundly to those fields; but Cameron knew his interpretive papers were successive approximations to the truth, not final answers, and right to the end he continued rethinking his vision of the cosmos. Appropriately, at the end of his life, Cameron was awarded the Hans A. Bethe Prize of the American Physical Society for his work in astrophysics and nuclear physics. He lived long enough to learn of this honour, but not to attend the prize ceremony, which was scheduled for the society's April 2006 meeting. Bethe, who wrote the classic 1939 paper on hydrogen burning in stars, had been the first to put his foot on the path that Cameron trod for so many years. ■

John Wood

John Wood is recently retired from the Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. e-mail: jawood@alum.mit.edu

NASA

BRIEF COMMUNICATIONS

No force limit on greyhound sprint speed

Unlike human athletes, these dogs do not need to slow down when racing round a tight bend.

Maximum running speed is constrained by the speed at which the limbs can be swung forwards and backwards, and by the force they can withstand while in contact with the ground. Humans sprinting around banked bends change the duration of foot contact to spread the time over which the load is applied, thereby keeping the force on their legs constant¹. We show here that, on entering a tight bend, greyhounds do not change their foot-contact timings, and so have to withstand a 65% increase in limb forces. This supports the idea that greyhounds power locomotion by torque about the hips, so — just as in cycling humans — the muscles that provide the power are mechanically divorced from the structures that support weight.

Sprinting around a bend increases effective body weight, as body mass experiences both gravity and centripetal acceleration (Fig. 1a). Human athletes respond to this by increasing the proportion of time per stride that each foot spends on the ground (the 'duty factor'). As swing time of the foot and its angle during

contact (stance angle) are constrained², this results in a reduction in speed¹. Athletes running on inside lanes, where bends are tighter, are at a disadvantage — see, for instance, the results from the 2004 World Indoor Championships 200-metre race³; the bias is so extreme that the indoor event has now been abandoned by the International Association of Athletics Federations.

To investigate the biomechanics of bend-running in greyhounds, we determined speed and footfall timings for 17 greyhounds from high-speed video recordings (250 frames per second) as they came to the end of the first straight (after more than 11 strides, 40 m), and then at the apex of the subsequent bend, which had a radius of 22.4 m (for movies, see supplementary information). Another 23 greyhounds, of similar ability to the first 17 dogs, were monitored as they started the second straight, directly after the bend. The dogs were undergoing pre-competition time trials, racing in small groups (1–3 individuals) around an arena composed of two straights and two semicircular bends of minimal banking. A



Dogged approach: greyhounds pound the ground around a bend.

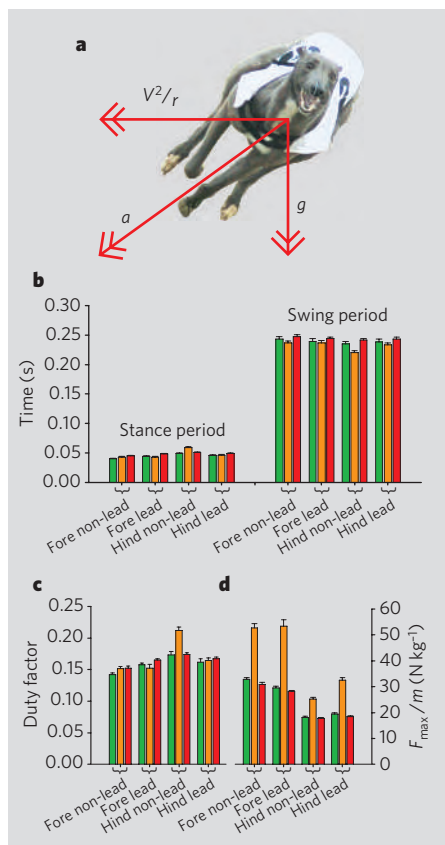


Figure 1 | Mechanics of bend-running in greyhounds. **a**, Mean acceleration vectors (drawn to scale) opposed by limb forces for greyhounds sprinting around a bend, where g is the acceleration due to gravity, V^2/r is the centripetal acceleration, and a is the resultant acceleration. **b**, Stance period (duration of foot contact with the ground) and swing period (the remainder of the stride) for individual limbs. Bars: green, first straight; orange, bend; red, second straight. **c**, Proportion of each stride spent by foot on the ground (duty factor) for individual limbs. **d**, Derived peak force on limbs estimated for greyhounds sprinting on the first straight and around a bend of radius 22.4 m (means ± 1 s.e., $n = 17$), where F_{max} is the maximum reaction force on each limb and m is body mass. Parameters for dogs of similar ability ($n = 23$) at the beginning of the second straight were similar to those for dogs on the first straight. Greyhounds experience substantially higher peak limb forces when running around the bend.

mechanical 'hare' on the outside of the track acted as a lure and was speed-controlled to elicit near-maximal performance.

Speed on the first straight ($V = 16.3 \pm 0.3$ m s⁻¹; mean ± 1 s.e. throughout) was significantly lower than on the bend (17.6 ± 0.2 m s⁻¹); however, net fore-aft acceleration was negligible and so these speeds are near-maximal (track sprint records require averages of 16–17 m s⁻¹). The mean speed for the second group of dogs was slightly lower at the start of the second straight ($V = 15.2 \pm 0.08$ m s⁻¹). The centripetal acceleration requirement (V^2/r) on the bend was high, resulting in an increase in effective weight of $71.0 \pm 2.3\%$.

Changes in the dogs' foot-contact timings (Fig. 1b) were mostly insignificant. Unlike humans, greyhounds do not compensate for the increased mean-force requirements of bend running by increasing the duty factor (Fig. 1c). Estimates of peak limb forces calculated from foot-contact timings^{4,5} (Fig. 1d, and see supplementary information) show that, compared with straight running before the bend, all four limbs experience a large increase in peak force during bend running (by $64.5 \pm 4.3\%$ on average for each leg). We conclude that the peak force on the legs does not constrain sprint speed in greyhounds during straight running under race conditions.

We attribute the different constraints on top speed in greyhounds and humans to their different mechanisms for producing power and supporting weight. In humans, the muscles that power sprinting are loaded by weight-induced compression forces along the leg. In contrast, the dogs power locomotion by torque

about the hips^{6–8} and by back extension; weight support is biased towards the forelimbs⁸. This mechanism is characteristic of cursorial quadrupeds. It is associated with muscular hip retractors and with forelimbs that are dominated by bone, tendon and highly pennate muscles, which act almost like passive springs^{9,10} and are capable of opposing considerable weight-induced forces. This means that the muscles that power greyhounds are virtually independent of weight support and so are not affected by an increase in effective weight.

James R. Usherwood*, Alan M. Wilson*†

*Structure and Motion Laboratory, The Royal Veterinary College, North Mymms, Hertfordshire AL9 7TA, UK

e-mail: jusherwood@rvc.ac.uk

†Centre for Human Performance, University College London, Stanmore, Middlesex HA7 4LP, UK

- Greene, P. R. *J. Biomech.* **20**, 667–680 (1987).
- Weyand, P. G., Sternlight, D. B., Bellizzi, M. J. & Wright, S. *J. Appl. Physiol.* **89**, 1991–1999 (2000).
- International Association of Athletics Federations www.iaaf.org/news/kind=4/newsId=24423.html (2005).
- Alexander, R. McN., Maloiy, G. M. O., Hunter, B., Jayes, A. S. & Nturi, J. *J. Zool. Lond.* **189**, 135–144 (1979).
- Witte, T. H., Knill, K. & Wilson, A. M. *J. Exp. Biol.* **207**, 3639–3648 (2004).
- Gray, J. *Animal Locomotion* (Weidenfeld & Nicolson, London, 1968).
- Alexander, R. McN. *J. Zool. Lond.* **173**, 549–573 (1974).
- Lee, D. V., Bertram, J. E. A. & Todhunter, R. J. *J. Exp. Biol.* **202**, 3565–3573 (1999).
- Gregersen, C. S., Siverton, N. A. & Carrier, D. R. *J. Exp. Biol.* **201**, 3197–3210 (1998).
- McGuigan, M. P. & Wilson, A. M. *J. Exp. Biol.* **206**, 1325–1336 (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none. doi:10.1038/438753a

ASTROPHYSICS

Is a doomsday catastrophe likely?

The risk of a doomsday scenario in which high-energy physics experiments trigger the destruction of the Earth has been estimated to be minuscule¹. But this may give a false sense of security: the fact that the Earth has survived for so long does not necessarily mean that such disasters are unlikely, because observers are, by definition, in places that have avoided destruction. Here we derive a new upper bound of one per billion years (99.9% confidence level) for the exogenous terminal-catastrophe rate that is free of such selection bias, using calculations based on the relatively late formation time of Earth.

Fears that heavy-ion collisions at the Brookhaven Relativistic Heavy Ion Collider might initiate a catastrophic destruction of Earth have

focused on three possible scenarios: a transition to a lower vacuum state that propagates outwards from its source at the speed of light²; formation of a black hole or gravitational singularity that accretes ordinary matter²; or creation of a stable 'strangelet' that accretes ordinary matter and converts it to strange matter³. A careful study¹ concluded that these hypothetical scenarios are overwhelmingly more likely to be triggered by natural high-energy astrophysical events, such as cosmic-ray collisions, than by the Brookhaven collider.

Given that life on Earth has survived for nearly 4 billion years (4 Gyr), it might be assumed that natural catastrophic events are extremely rare. Unfortunately, this argument is flawed because it fails to take into account an observation-selection effect^{4,5}, whereby observers are precluded from noting anything other than that their own species has survived up to the point when the observation is made. If it takes at least 4.6 Gyr for intelligent observers to arise, then the mere observation that Earth has survived for this duration cannot even give us grounds for rejecting with 99% confidence the hypothesis that the average cosmic neighbourhood is typically sterilized, say, every 1,000 years. The observation-selection effect guarantees that we would find ourselves in a lucky situation, no matter how frequent the sterilization events.

Figure 1 indicates how we derive an upper bound on the cosmic catastrophe frequency τ^{-1} that is free from such observer-selection bias. The idea is that if catastrophes were very frequent, then almost all intelligent civilizations would have arisen much earlier than ours. Using data on planet-formation rates⁶, the distribution of birth dates for intelligent species

can be calculated under different assumptions about the rate of cosmic sterilization. Combining this with information about our own temporal location enables us to conclude that the cosmic sterilization rate for a habitable planet is, at most, of the order of 1 per 1.1 Gyr at 99.9% confidence. Taking into account the fact that no other planets in our Solar System have yet been converted to black holes or strange matter^{1–3} further tightens our constraints on black hole and strangelet disasters. (For details, see supplementary information.)

This bound does not apply in general to disasters that become possible only after certain technologies have been developed — for example, nuclear annihilation or extinction through engineered microorganisms — so we still have plenty to worry about. However, our bound does apply to exogenous catastrophes (for example, those that are spontaneous or triggered by cosmic rays) whose frequency is uncorrelated with human activities, as long as they cause permanent sterilization. Using the results of the Brookhaven analysis¹, the bound also implies that the risk from present-day particle accelerators is reassuringly small: say, less than 10^{-12} per year.

Max Tegmark*, Nick Bostrom†

*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

e-mail: tegmark@mit.edu

†Future of Humanity Institute, Faculty of Philosophy, University of Oxford, Oxford OX1 4JJ, UK

- Jaffe, R. L., Busza, W., Sandweiss, J. & Wilczek, F. *Rev. Mod. Phys.* **72**, 1125–1140 (2000).
- Hut, P. & Rees, M. *J. Nature* **302**, 508–509 (1983).
- Dar, A. & De Rújula, A. *Phys. Lett. B* **470**, 142–148 (1999).
- Carter, B. in *IAU Symposium 63* (ed. Longair, M. S.) 291–298 (Reidel, Dordrecht, 1974).
- Bostrom, N. *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge, New York, 2002).
- Lineveaver, C. H., Fenner, Y. & Gibson, B. K. *Science* **203**, 59–62 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none. doi:10.1038/438754a

CORRIGENDUM

Avian flu: Isolation of drug-resistant H5N1 virus

Q. Mai Le, Maki Kiso, Kazuhiko Someya, Yuko T. Sakai, T. Hien Nguyen, Khan H. L. Nguyen, N. Dinh Pham, Ha H. Ngyen, Shinya Yamada, Yukiko Muramoto, Taisuke Horimoto, Ayato Takada, Hideo Goto, Takashi Suzuki, Yasuo Suzuki, Yoshihiro Kawaoka
Nature **437**, 1108 (2005)

We omitted the accession numbers for the sequences of the A/Hanoi/30408/2005 clones, which are registered in the DNA Data Bank of Japan. These are:

AB239125 20051020120345.25409 for the haemagglutinin gene in clone 9; and
AB239126 20051020122743.63420 for the neuraminidase gene in clone 7.

doi:10.1038/438754b

BRIEF COMMUNICATIONS ARISING online
♦ www.nature.com/bca see Nature contents.

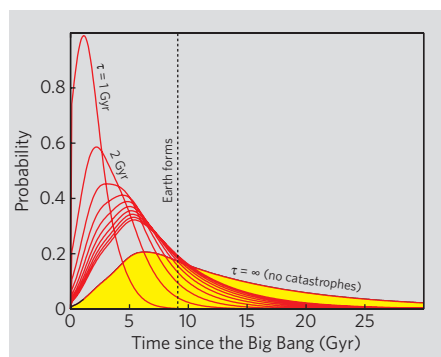


Figure 1 | The catastrophe timescale cannot be very short. The probability distribution is shown for observed planet-formation times, assuming catastrophe timescales, τ , of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 Gyr and infinity (shaded yellow), respectively (from left to right). The probability of observing a formation time ≥ 9.1 Gyr for Earth (area to the right of the dotted line) drops below 0.001 for $\tau < 1.1$ Gyr.

PLANETARY SCIENCE

Are there active glaciers on Mars?

Arising from: J. W. Head *et al.* *Nature* **434**, 346–351 (2005)

Head *et al.*¹ interpret spectacular images from the Mars Express high-resolution stereo camera as evidence of geologically recent rock glaciers in Tharsis and of a piedmont ('hourglass') glacier at the base of a 3-km-high massif east of Hellas. They attribute growth of the low-latitude glaciers to snowfall during periods of increased spin-axis obliquity. The age of the hourglass glacier, considered to be inactive and slowly shrinking beneath a debris cover in the absence of modern snowfall, is estimated to be more than 40 Myr. Although we agree that the maximum glacier extent was climatically controlled, we find evidence in the images to support local augmentation of accumulation from snowfall through a mechanism that does not require climate change on Mars.

Head *et al.*¹ identify an accumulation area for the hourglass glacier in an 'alcove' above its upper crater (Fig. 1a). However, the arcuate banded ice that originated from the canyon draining most of the alcove is topographically lower than, and appears to be pinched out by, flow from either side (<http://esamultimedia.esa.int/images/marsexpress/180-170305-0451-6-an-01-Hourglass.jpg>). Hence, the greatest ice flux must have come from small areas at the foot of the adjacent slopes and not from the larger area drained by the canyon, as would be expected for a snowfall-fed glacier. This suggests that the most recent source of ice was not associated with snowfall in the accumulation area, even though earlier it may have been.

What is the source of the most recent ice? Insight may come from another apparently ice-filled double crater (Fig. 1b) that lies on the flat plain only 35 km away and at the same elevation as the distal crater, according to Mars Orbiter Laser Altimeter data (<http://ftpwww.gsfc.nasa.gov/tharsis/mola.html>). Despite the absence of potential accumulation areas, similarly banded and presumably debris-covered icy flows still fill the craters and point to a source within each crater. The observation that these glaciers lack distinct accumulation areas, together with the absence of glaciers in many comparable locations elsewhere on the surrounding plain, argues against a solely climatic origin.

These observations are consistent with the glaciers having been formed by groundwater erupting from exposed aquifers, or from fractures in an icy regolith, and then freezing near the surface (*aufeis*). This idea was proposed by Garlick² for a palaeoglacier that originated from a fissure on Arsia Mons.

Where could such large amounts of ground-

water come from? Possibilities on Mars include melting permafrost and dewatering of hydrous compounds^{3–7} by heating due to increased geothermal gradients, magma intrusions or impacts. The long-term stability of shallow groundwater and ice at the latitude of the hourglass craters is debatable. But geologically recent gullies heading part way up slopes have been interpreted as evidence for shallow deposits (at depths of less than 0.5 km) of groundwater or ice at similar latitudes, including locations nearby⁸. Moreover, *aufeis* feeding of glaciers is not directly related to climate fluctuations, raising the possibility that some of the glaciers identified by Head *et al.*¹ remain active.

We find support for the *aufeis* hypothesis in a Mars Orbital Camera high-resolution image of the hourglass glacier (Fig. 1c), which shows a band of light-coloured, crevassed ice that is within about 250 m of the uphill edge of the glacier. It seems to be flowing from the slope at the southern edge of the crater to the west of the canyon. Downslope, the crevassed ice becomes progressively covered by dark debris. The Mars Orbital Camera image cited by Head *et al.*¹ shows only the distal, lightly cratered, inactive part of the glacier. Because ice on Mars sublimates rapidly until it is covered, the exposed and crevassed ice in the proximal glacier must be active. The presence of relatively bare, subliming ice and the absence of an accumulation area raise the possibility that water, or ice, is erupting on the martian surface today.

We conclude that martian glaciers could be fed from above or below. This could be by snowfall during climatically favourable times, as argued by Head *et al.*¹, or by extrusion of groundwater, as we argue here. These two mechanisms are not mutually exclusive and their relative contributions may differ spatially and vary temporally: during climates that support snowfall, both may operate; at other times, diminished glaciers may still be supported by *aufeis*. We consider the relative contribution of snowfall and extrusion mechanisms to the growth and extent of glaciers as an outstanding question in martian glaciology. Alan R. Gillespie, David R. Montgomery, Amit Mushkin

Quaternary Research Center and Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195-1310, USA
e-mail: alan@rad.ess.washington.edu

- Head, J. W. *et al.* *Nature* **434**, 346–351 (2005).
- Garlick, G. D. *Geol. Soc. Am. Cordilleran Section abstr.* **20**, 162 (1988).
- Sharp, R. P. *J. Geophys. Res.* **78**, 4073–4083 (1973).

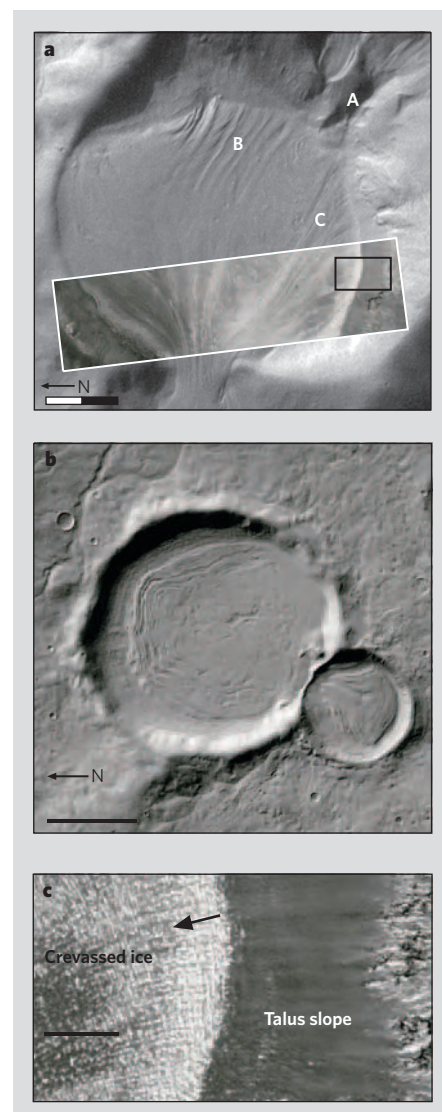


Figure 1 | Images interpreted to show *aufeis* on Mars (257.3°W, –39.5°S). **a**, High-resolution stereo camera (HRSC) image (about 20 m per pixel) of the upper basin of the 'hourglass' glacier and the mouth of the canyon (location A) draining the largest part of the proposed accumulation area (not shown); inferred *aufeis* source areas are near locations marked B and C. Images are taken from Mars Express orbit 451. The overlay (white outline) is part of higher-resolution (4.2 m) Mars Orbital Camera (MOC) image, M1102128. Black box, area enlarged in **c**. **b**, A second 'hourglass' glacier from the same HRSC image, about 35 km north of the area shown in **a**. **c**, Enlargement of the area in a black box in the MOC overlay in **a**. The reticulate pattern is interpreted as a crevasse field below the talus slope. Arrow showing flow direction spans the light-coloured, relatively bare ice that becomes increasingly covered farther north. Scale bars: **a**, 2 km; **b**, 5 km; **c**, 300 m.

4. Luchitta, B. K. *Icarus* **45**, 264–303 (1981).
5. Komatsu, G. *et al.* *Lunar Planet. Sci. XXXI* abstr. 1434 (CD-ROM) (Lunar and Planetary House, Houston, 2000).
6. Kargel, J. *Mars — A Warmer, Wetter Planet* (Springer-Praxis, Chichester, UK, 2004).
7. Montgomery, D. R. & Gillespie, A. *Geology* **33**, 625–628 (2005).
8. Malin, M. C. & Edgett, K. S. *J. Geophys. Res.* **106**, 23429–23570 (2001).

doi:10.1038/nature04357

PLANETARY SCIENCE

Head *et al.* reply

Replying to: A. R. Gillespie, D. R. Montgomery & A. Mushkin *Nature* **438**, doi:10.1038/nature04357 (2005)

Gillespie *et al.*¹ concur with our interpretation that certain lobate equatorial and mid-latitude features on Mars are due to debris-covered glaciers formed largely during past periods of increased spin-axis obliquity, when climate regimes favoured snow and ice accumulation and glacial flow². They suggest that the 'hourglass' deposit, dated at more than 40 Myr old³, could be active today owing to an additional mechanism that supports "local augmentation of accumulation from snowfall" without climate change on Mars. This mechanism requires the present, or very recent, release of groundwater to the surface to form *aufeis* (groundwater-fed 'glaciers') where the groundwater is generated by dewatering of hydrous compounds or melting by magmatic or impact-generated heat. We assess whether this suggestion applies to the deposits in question — it was previously proposed for much older deposits in other areas of Mars^{3,4}. We make particular reference to the key relationships in the accumulation zones.

Glacial accumulation zones are areas characterized by a yearly net addition of ice, and are separate from ablation zones, which undergo a yearly net loss of ice. In active valley glaciers and in the source regions for broader piedmont glaciers, the accumulation zones are commonly centred on steep-sided alcoves that provide broad traps for wind-blown snow and

lie in shadow, producing colder temperatures. The accumulation zone for the hourglass deposit on Mars was thought to be the entire alcove region east and southeast of the hourglass deposit² (Fig. 2a, b, d of ref. 2). Snow and ice accumulating in this entire alcove region probably formed distinct glaciers that moved downslope and converged to form multiple lobes, the remnants of which are seen in the upper part of the hourglass.

Changes in ice velocity, as might occur with changes in ice thickness related to deepening of bedrock topography, compressed and folded the ice, eventually merging all lobes together, then to flow through the neck of the hourglass. When climatic conditions changed, the distribution of snowfall and snow accumulation probably also changed, resulting in the cessation of ice accumulation and flow; the ice and snow deposits disappeared, and dry mass-wasting dominated, producing the talus piles superposed on top of the glacial deposits at the base of the slope (Fig. 2a–d of ref. 2, and Fig. 1a, c of ref. 1).

We therefore interpret the configuration and morphology of the lobes, as well as the abrupt contact between the base of the slope and the glacial deposits, to be a natural consequence of both the shape of the initial accumulation zone and its postglacial evolution. The candidate *aufeis* in the high-resolution image (Fig. 1c of ref. 1) appears

morphologically similar to the floor deposits across the entire frame; we interpret its relative brightness to be due to a dust cover that was formed on the margin of the encroaching talus slope.

Nearby craters with similar scales, hourglass-like shapes, and lobate flow-like deposits (Fig. 1b of ref. 1) can also form from local crater-wall accumulation zones. In these cases, the crater interior walls are optimum locations for snow and ice accumulation, resulting in glacier-like, lobate flow down the walls and out onto the crater floor, a phenomenon that has been well documented in mid-latitude crater interiors^{5,6}. Alcoves along crater walls create an environment favourable to the collection of snow and ice, and may provide steep cliffs that act as sources for debris that, together with ice sublimation, create debris-covered glaciers, much like those that occur in the Mars-like Antarctic Dry Valleys (Fig. 3a, b of ref. 2).

On the basis of these considerations, groundwater-fed 'glaciers' do not seem to be required in these locations. We agree that this mechanism should be investigated, particularly for deposits formed earlier in the history of Mars, when thermal gradients were such that groundwater may have been much closer to the surface.

J. W. Head*, G. Neukum, R. Jaumann, H. Hiesinger, E. Hauber, M. Carr, P. Masson, B. Foing, H. Hoffmann, M. Kreslavsky, S. Werner, S. Milkovich, S. van Gasselt, The HRSC Co-Investigator Team

*Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA
e-mail: james_head_III@brown.edu

1. Gillespie, A. R., Montgomery, D. R. & Mushkin, A. *Nature* **438**, doi:10.1038/nature04357 (2005).
2. Head, J. W. *et al.* *Nature* **434**, 346–351 (2005).
3. Carr, M. J. *J. Geophys. Res.* **100**, 7479–7507 (1995).
4. Lucchitta, B. K. *J. Geophys. Res.* **89**, suppl. B 409–418 (1984).
5. Howard, A. *Lunar Planet. Sci. XXXIV* abstr. 1065 (2003).
6. Marchant, D. R. & Head, J. W. *6th Int. Mars Conf.* abstr. 3991 (2003).

doi:10.1038/nature04358

NEWS & VIEWS

PLANETARY SCIENCE

Huygens rediscovers Titan

Tobias Owen

The first analyses of data sent by the Huygens probe from Saturn's largest moon Titan are flooding in. They paint a picture of a 'Peter Pan' world — potentially like Earth, but with its development frozen at an early stage.

Ever since the two Voyager spacecraft passed the Solar System's sixth planet in 1980 and 1981, Saturn — with its beautiful rings and retinue of more than 30 satellites cocooned in a complex, pulsating magnetosphere — has insistently called on us to return. The joint NASA and European Space Agency (ESA) Cassini–Huygens mission, launched from Cape Canaveral on 15 October 1997, was the answer to that call. In this issue, an overview¹ is given of the descent of the Huygens probe through the atmosphere of the largest saturnian moon, Titan, and its subsequent landing on the satellite's surface. The first results^{2–7} from the six instruments on board are also presented. These data, even at such an early stage of analysis, are highly enlightening — and are generating exciting questions.

To reach this stage, the Cassini–Huygens mission had first to overcome numerous political and technical challenges. The hardware that was needed to investigate Saturn — an orbiter and at least one atmospheric probe — was beyond the means of any one space agency, calling for a new form of international partnership. So NASA constructed the Cassini orbiter (named after the astronomer Giovanni Domenico Cassini, who discovered four of Saturn's moons), which would carry a range of instruments to investigate the entire Saturn system. ESA, meanwhile, built the Huygens probe (named after Christiaan Huygens, the Dutch physicist and astronomer who discovered Titan) that would enter Titan's atmosphere and descend to its surface. Scientists and engineers from both sides participated in all phases of the project, contributing instruments to both orbiter and probe.

This bold initiative has proved brilliantly successful: after a seven-year journey and insertion into orbit around Saturn, Cassini released Huygens on 25 December 2004. The probe entered Titan's atmosphere on 14 January 2005 and successively deployed its parachutes, taking 2 hours, 27 minutes to descend to the satellite's surface. Once there, it transmitted data via Cassini and a network of

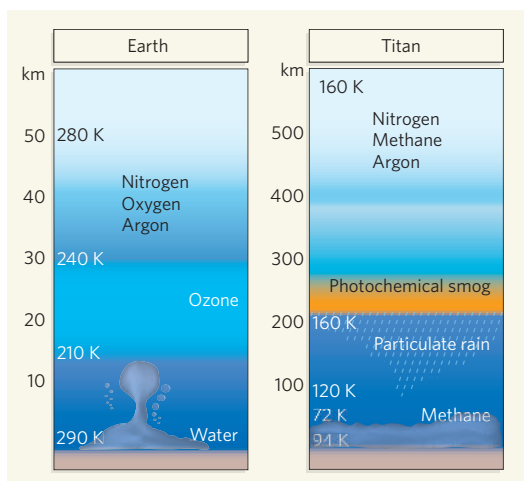


Figure 1 | Comparison between the atmospheres of Earth and Titan. The descent of the Huygens probe^{1–7} has allowed the first detailed study of the atmosphere of Saturn's moon Titan, revealing startling parallels — and stark contrasts — with that of Earth. Both atmospheres are nitrogen-dominated, but the low temperature of Titan means that the carbon-carrying gas in its atmosphere is methane (1.6% of the total) rather than carbon dioxide (present at only 345 parts per million). Photochemical reactions involving this methane produce a smog at middle altitudes, and an organic rain of methane and nitrogen-containing aerosols falls steadily onto the satellite's surface, creating an Earth-like terrain of extended river networks. Radiogenic argon (⁴⁰Ar), which makes up 1% of Earth's atmosphere, is in short supply on Titan (just 43 parts per million). The still smaller amount of primordial argon (³⁶Ar) suggests that the nitrogen in the atmosphere must have arrived in the form of compounds such as ammonia, rather than as molecular nitrogen.

antennas around the Earth to the mission's command centre in Darmstadt, Germany, for a further 69 minutes. These were the first signals from the icy surface of a satellite that lies ten times farther from the Sun than does Earth. The reaction of the waiting scientists and engineers is best reflected in the words Huygens himself used as he attempted to imagine Galileo's feelings on discovering Jupiter's satellites: "No small rapture."

But why Titan? Because of its immense distance from the Sun, Titan's development was frozen at a very early stage — where it will remain until the Sun develops into a red giant

star and melts it. Larger than the planet Mercury, Titan is a world massive enough and cold enough to have a nitrogen-dominated atmosphere ten times thicker than our own (Fig. 1). Its extremely low temperature keeps water frozen, so that even water vapour is missing from the atmosphere. In contrast, on the warmer inner planets Mars, Earth and Venus, where water vapour is active, carbon compounds become quickly oxidized.

Without water as a source of oxygen, primitive, hydrogen-rich conditions have existed on Titan for billions of years, as signalled by the fact that the dominant carbon-carrying gas is not carbon dioxide, but methane (CH₄). In this atmosphere, photochemical reactions produce thick layers of organic smog that prevent Titan's surface from being viewed at visible wavelengths. The surface temperature of Titan measured by the Huygens Atmospheric Structure Instrument (HASI)² is just 94 K, or -179°C ; this would allow the existence of lakes and rivers of liquid methane. Yet this exotic, highly flammable world may offer illuminating insights into the hidden history of the early Earth.

The basic physical characteristics of Titan's atmosphere were determined by HASI and the Doppler Wind Experiment during the Huygens probe's descent. The latter instrument established that the winds at lower altitudes in Titan's atmosphere blow on average in the direction in which Titan is rotating³, reaching a maximum of 120 metres per second (430 km h^{-1}) at an altitude of about 120 km. This 'super-rotation', also seen in Venus's atmosphere, confirms both Earth-based observations of Titan and theoretical models⁸. HASI found that the winds at Titan's surface are, in contrast, very weak, with speeds of around 1 metre per second (3.6 km h^{-1}) or less³. The challenge raised by this observation is to find out whether such light winds can account for the observed wind-induced features on Titan's surface, or whether stronger gusts are required.

The Huygens probe's Aerosol Collector and Pyrolyser (ACP)⁴ captured and heated aerosols

NASA

— suspended liquid and solid particles — in Titan's atmosphere during descent, sending the effluent to the Gas Chromatograph Mass Spectrometer (GCMS)⁵ for analysis. Preliminary evidence indicates the presence of nitrogen-containing organic compounds, which may include amino, imino and nitrile groups (further analysis is being carried out to determine the precise compounds present). These aerosols must fall steadily as a kind of organic rain on Titan's surface, producing a global layer with a potential thickness of 1 kilometre or more⁹. This means that the information from the ACP can be used to predict what additional compounds might be formed on the surface. Relatively recent depositions probably cause the dark patches seen on Titan's surface in infrared images¹⁰. The process removes atmospheric methane; indeed, methane would disappear completely from Titan's atmosphere in just 10 million to 20 million years without some (as yet undiscovered) source to replace it, either continuously or episodically.

The results from several of Huygens' instruments imply that methane is involved in a phase-change cycle on Titan similar to that of water on Earth. Both the GCMS and the Descent Imager/Spectral Radiometer (DISR)⁶ detected a methane haze in the lower regions of the atmosphere, and the Surface-Science Package⁷ found the ground at the landing site to have the consistency of damp sand — probably a mixture of ice chips, precipitated aerosols and some liquid. The liquid was identified as condensed methane by a sudden surge in the methane signal detected by the GCMS after the probe had landed. This higher signal level lasted for the remaining 69 minutes of communication with the Cassini orbiter⁵.

It would therefore seem, to update Professor Higgins's exercise in elocution, that the titanian rain is mainly methane. But despite the fact that exotic materials produced it, the landscape near the probe is remarkably Earth-like: stunning images from the DISR reveal beautifully defined channel systems incised into the surrounding terrain⁶ (Fig. 2). These river networks must have been cut into the bedrock of thick ice by some combination of methane rain and subsurface springs.

Another feature that Titan shares with Earth

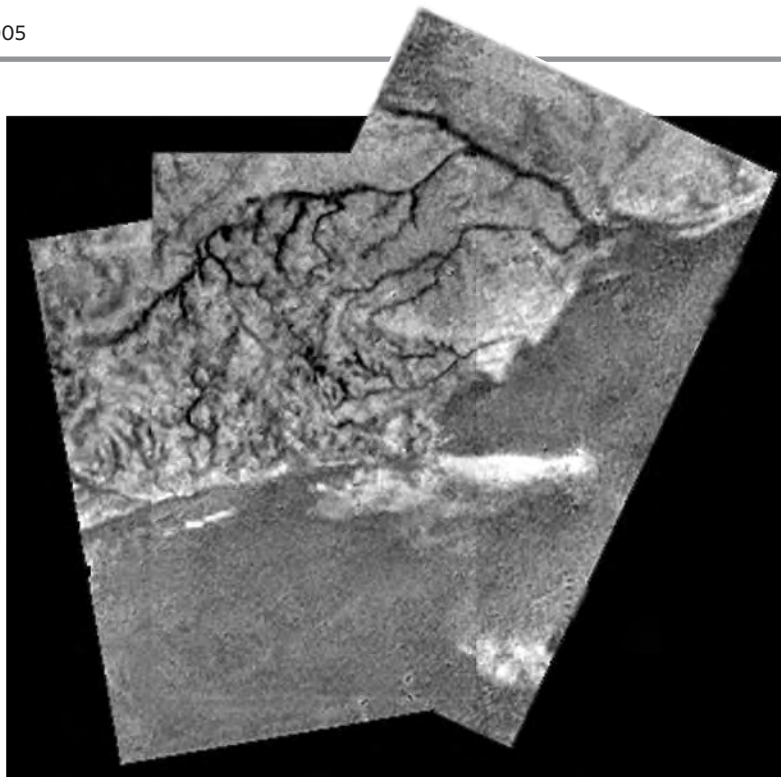


Figure 2 | Rivers of methane. A mosaic of three frames of Titan's surface, taken by the Huygens probe from an altitude of 16 km, showing a system of converging river channels.

is the abundant nitrogen in its atmosphere. Relative to its abundance in Earth's atmosphere, however, the lighter nitrogen isotope ¹⁴N is depleted on Titan, indicating that roughly five times the present amount of atmosphere has escaped from Titan since its formation⁵. (Escape into space is the only plausible way to selectively remove one isotope, and has been observed on Mars¹¹.) Carbon, which is present in methane, does not show such a large anomalous isotopic ratio — a further indication that methane must be continually replenished on Titan.

One possible pathway for this methane replenishment is cryovolcanism, or a similar mechanism by which material would be exuded by Titan's interior.

This possibility is supported by the discovery in Titan's atmosphere of the argon isotope ⁴⁰Ar, which originates solely from the decay of radioactive potassium; this potassium must exist in rocks predominantly below the satellite's ice-water mantle. In contrast, the atmosphere contains only a trace of the primordial isotope ³⁶Ar. This fact provides us with an important clue to the conditions under which Titan formed: it indicates that nitrogen originally arrived on Titan as a mixture of compounds such as ammonia (as indeed it came to Earth). Simple molecular nitrogen, N₂, is trapped only in icy planetesimals of the kind that created Titan at temperatures below 45 K, at which a large fraction of the ambient ³⁶Ar would also be trapped. Compounds such as ammonia can be trapped at much higher temperatures, without

trapping primordial argon.

But it is not just the primordial noble gases that are missing in Titan's atmosphere — for any world to have a thick nitrogen atmosphere, an amount of carbon 4 to 20 times greater than that actually found must be hidden somewhere. (The range is given by the proportions found elsewhere in nature, the lower figure being given by the ratio of abundances in the Sun, and the higher figure by that found in comets and the atmospheres of the inner planets^{12,13}.) On Earth, the missing carbon is found in enormous deposits of carbonate rocks in the crust. On Titan, some might exist as ancient deposits of aerosols, buried under layers of ice that have successively resurfaced the satellite, hiding the

impact craters that once dominated the landscape. But the need for a replenishing source of methane suggests that vast deposits of carbon may still be sequestered deep inside Titan, perhaps in the form of methane that was made in the satellite's mantle eons ago. This methane could now be held in cage-like structures of water molecules (a form known as a clathrate hydrate) at the bottom of an ocean hypothesized to lie beneath Titan's crust. Alternatively, a reservoir of carbon in its original form (organic compounds, carbon dioxide and grains) that is still being converted to methane might exist far below the surface⁵.

The mystery of the missing methane is just one of many to be solved. Titan will have much to tell us even after the rich harvest of the Huygens data has been analysed. The region where the probe landed was named 'Antilia', after a mythical island once thought to lie between Europe and the Americas. That symbol of the international nature of the Cassini-Huygens endeavour should serve as an inspiration for the next mission to set sail for this fiercely frozen echo of the early Earth.

Tobias Owen is at the Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA.
e-mail: owen@ifa.hawaii.edu

1. Lebreton, J.-P. *et al.* *Nature* **438**, 758–764 (2005).
2. Fulchignoni, M. *et al.* *Nature* **438**, 785–791 (2005).
3. Bird, M. K. *et al.* *Nature* **438**, 800–802 (2005).
4. Israël, G. *et al.* *Nature* **438**, 796–799 (2005).
5. Niemann, H. B. *et al.* *Nature* **438**, 779–784 (2005).
6. Tomasko, M. G. *et al.* *Nature* **438**, 765–778 (2005).
7. Zarnecki, J. C. *et al.* *Nature* **438**, 792–795 (2005).
8. Hord, F. *et al.* *Icarus* **117**, 358–374 (1995).
9. Strobel, D. F. *Planet. Space Sci.* **30**, 833–838 (1982).
10. Porco, C. C. *et al.* *Nature* **434**, 159–168 (2005).
11. McElroy, M. B. *et al.* *J. Geophys. Res.* **82**, 4379–4388 (1977).
12. Owen, T. & Bar-Nun, A. *Icarus* **116**, 215–226 (1995).
13. Grevesse, N. *et al.* *EAS Publ. Ser.* **9**, 1 (2005).

"The brightest of Saturn's [satellites], it chanc'd to be my lot, with a telescope not above 12 foot long, to have the first light of in the year 1655. The rest we may thank the industrious Cassini for..."
— Christiaan Huygens, 1685

ARTICLES

An overview of the descent and landing of the Huygens probe on Titan

Jean-Pierre Lebreton¹, Olivier Witasse¹, Claudio Sollazzo³, Thierry Blancquaert², Patrice Couzin⁴, Anne-Marie Schipper⁴, Jeremy B. Jones⁵, Dennis L. Matson⁵, Leonid I. Gurvits⁶, David H. Atkinson⁷, Bobby Kazeminejad⁸ & Miguel Pérez-Ayúcar¹

Titan, Saturn's largest moon, is the only Solar System planetary body other than Earth with a thick nitrogen atmosphere. The Voyager spacecraft confirmed that methane was the second-most abundant atmospheric constituent in Titan's atmosphere, and revealed a rich organic chemistry, but its cameras could not see through the thick organic haze. After a seven-year interplanetary journey on board the Cassini orbiter, the Huygens probe was released on 25 December 2004. It reached the upper layer of Titan's atmosphere on 14 January and landed softly after a parachute descent of almost 2.5 hours. Here we report an overview of the Huygens mission, which enabled studies of the atmosphere and surface, including *in situ* sampling of the organic chemistry, and revealed an Earth-like landscape. The probe descended over the boundary between a bright icy terrain eroded by fluvial activity—probably due to methane—and a darker area that looked like a river- or lake-bed. Post-landing images showed centimetre-sized surface details.

Titan is the second-largest moon in the Solar System, after Jupiter's Ganymede, and is assumed to have formed in the Saturn subnebula about 4.5 billion years ago. One of its great mysteries is the origin of the methane in the atmosphere. With a lifetime of just 20 million years, methane must be regularly resupplied to the atmosphere to be as abundant as it is today. The surface of Titan remained hidden to the Voyager cameras, which led to speculation on its appearance and processes. The surface pressure on Titan is about 1.5 times that on Earth and the surface temperature is about minus 180 °C. At such a low temperature, it was postulated that liquid methane might be present on Titan's surface or in underground reservoirs. Although the images returned by the Voyager spacecraft were featureless, the richness of the detected organic compounds confirmed that Titan was indeed worthy of being revisited and explored in detail.

The distinct orange appearance of Titan's atmosphere, as observed by the Voyagers in the early 1980s, comes from the methane-induced organic chemistry. Complex hydrocarbons and carbon-nitrogen-based compounds form high in the atmosphere, which is irradiated by solar ultraviolet rays and bombarded by energetic particles from Saturn's space environment. Methane converts to ethane, acetylene, ethylene, and so on, and when combined with nitrogen forms hydrogen cyanide and more complex nitrogen-bearing carbon and hydrocarbon compounds. These organic compounds float slowly downward in the atmosphere, condense in the stratosphere, and form the aerosols that give the well-known orange colour to Titan's hazy atmosphere. The aerosols eventually rain to the surface, where they accumulate.

Images of Titan's surface at various resolutions were obtained by the Hubble Space Telescope^{1,2} and ground-based observatories^{3,4}. Early images of Titan's surface obtained by the Cassini orbiter⁵ were almost as baffling as those obtained from Earth. Bright and dark patches were clearly visible on the surface. Albedo patterns suggested a heterogeneous active surface, perhaps with some fluvial

processes. No direct evidence of surface liquid was found before the Huygens probe, although ground-based radar observations were interpreted as indicative of the presence of liquid surfaces⁶ near the equator.

The scientific objectives established for the Cassini-Huygens mission at Titan^{7,8} were to: (1) determine atmospheric composition; (2) investigate energy sources for atmospheric chemistry; (3) study aerosol properties and cloud physics; (4) measure winds and global temperatures; (5) determine properties of the surface and infer internal structure; and (6) investigate the upper atmosphere and ionosphere. The Huygens probe performed detailed *in situ* observations along the descent path and on the surface, while in mid-2004 the Cassini orbiter started to carry out the global mapping planned during its 45 Titan fly-bys. The Huygens probe's scientific payload includes six experiments: (1) HASI (Huygens Atmospheric Structure Instrument)⁹; (2) the DWE (Doppler Wind Experiment)¹⁰; (3) the ACP (Aerosol Collector and Pyrolyzer)¹¹; (4) the GCMS (Gas Chromatograph and Mass Spectrometer)¹²; (5) the SSP (Surface Science Package)¹³; (6) the DISR (Descent Imager and Spectral Radiometer)¹⁴. The main characteristics of the payload are given in the Supplementary Information. The payload accommodation is illustrated in Fig. 1.

The entire Huygens mission was designed to be carried out during a 2.5-hour descent through the atmosphere and possibly a few more minutes on the surface⁷. The probe was not guaranteed to survive its impact on what was unknown terrain. The coordinates of the predicted Huygens landing site were uncertain by several hundred kilometres because there were large uncertainties in how far the winds would carry the probe laterally during its descent under parachute.

After a seven-year interplanetary journey and two orbits around Saturn aboard the Cassini orbiter, the Huygens probe was released during the third orbit on 25 December 2004. On 14 January 2005 it

¹Research and Scientific Support Department, ²Scientific Project Department, ESA Science Directorate, ESTEC, 2200 AG, Noordwijk, the Netherlands. ³ESA Operations Directorate, ESOC, 64293 Darmstadt, Germany. ⁴Alcatel Alenia Space, BP 99, F-06156, Cannes-La Bocca, France. ⁵Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. ⁶Joint Institute for VLBI in Europe, PO Box 2, 7990 AA Dwingeloo, The Netherlands. ⁷University of Idaho, Department of Electrical and Computer Engineering, Moscow, Idaho 83844-1023, USA. ⁸Space Research Institute, Austrian Academy of Sciences, 8042 Graz, Austria.

reached the upper layers of Titan's atmosphere with a speed of 6 km s^{-1} . Analysis of the accelerometer measurements obtained during the entry produced an atmospheric structure profile from an altitude of 1,400 km down to 155 km (ref. 15). The Huygens probe revealed much structure above 200 km and the high-altitude densities were slightly more than predicted. The Huygens probe decelerated to 400 m s^{-1} in less than 5 min. The main parachute was then deployed at a speed of approximately mach 1.5 (400 m s^{-1}), at an altitude of about 155 km. Within one minute after the parachute deployment, the Huygens payload was fully operational and data were being transmitted to Cassini via two redundant radio-link channels¹⁶. After a descent of 2 h 28 min, the probe landed softly at 11:38:11 UTC (Coordinated Universal Time) with a speed slightly less than 5 m s^{-1} on a surface that appeared solid, with no apparent surface liquid, although some evaporated methane was detected soon after the impact. Cassini received data from the Huygens probe until 12:50 UTC (1 h 12 min after touchdown) when it passed below the probe's horizon. At that moment, the probe was still operating and broadcasting from the surface. One of the two carrier frequencies was received by an array of Earth-based radio telescopes¹⁷ (see the Supplementary Information), which provided wind measurements¹⁸.

Clear images of the surface were obtained below about 40 km altitude. The Huygens probe revealed an extraordinary world, resembling Earth in many respects, especially in meteorology, geomorphology and fluvial activity. The images show strong evidence for erosion due to liquid flows, possibly methane, on Titan. The probe trajectory carried it across a boundary between a bright, icy, rugged terrain and a darker flat area¹⁹. The Huygens probe landed in the dark area. The measured pressure and temperature profiles¹⁵ below 150 km are close to those expected on the basis of Voyager observations. The measured surface temperature and pressure at the

landing site were $\sim 93.7 \text{ K}$ and $\sim 1,470 \text{ mbar}$ respectively. At the landing site, the surface is relatively flat and solid. Reflectance spectra show that it is mostly composed of dirty water-ice¹⁹. Water-ice pebbles up to a few centimetres in diameter were scattered near the landing site. The Huygens Surface Science Package penetrometer found the surface here to be unconsolidated²⁰, with the consistency of loose wet sand.

Winds were found to blow predominantly in the direction of Titan's rotation¹⁸. West-east winds up to 450 km h^{-1} were detected above an altitude of 120 km. The winds decreased with decreasing altitude. An unexpected layer of high wind-shear was encountered between altitudes of 100 and 60 km. Perhaps unrelated but worth noting, an ionosphere-like layer produced by galactic cosmic rays was discovered at an altitude of about 60 km (ref. 15). The winds, with speeds of metres per second, reversed direction below 8 km (refs 18, 19). Haze was detected all the way down to the surface¹⁹, contrary to the predictions of pre-Huygens models. It was predicted that the atmosphere would be clear of haze in the lower stratosphere, below around 60 km. Fortunately, the haze was transparent enough for good images of the surface to be obtained below 40 km.

In situ composition measurements of Titan's atmosphere²¹ and of the aerosols²² below 150 km confirmed the presence of a complex organic chemistry in both the gas and the solid phase. Vertical profiles were obtained for the more abundant species. So far no new organic compound has been detected in the atmosphere, except that the presence of ^{40}Ar , already detected in the upper atmosphere by INMS²³, was confirmed. Primordial argon, ^{36}Ar , was detected, to our knowledge for the first time, but not xenon and krypton. The non-detection of these noble gases, a surprising finding, will fuel theories of the origin and evolution of Titan's atmosphere. The C, N and H/D isotopic ratios were measured. This will make it possible to constrain formation scenarios for Titan's atmosphere.

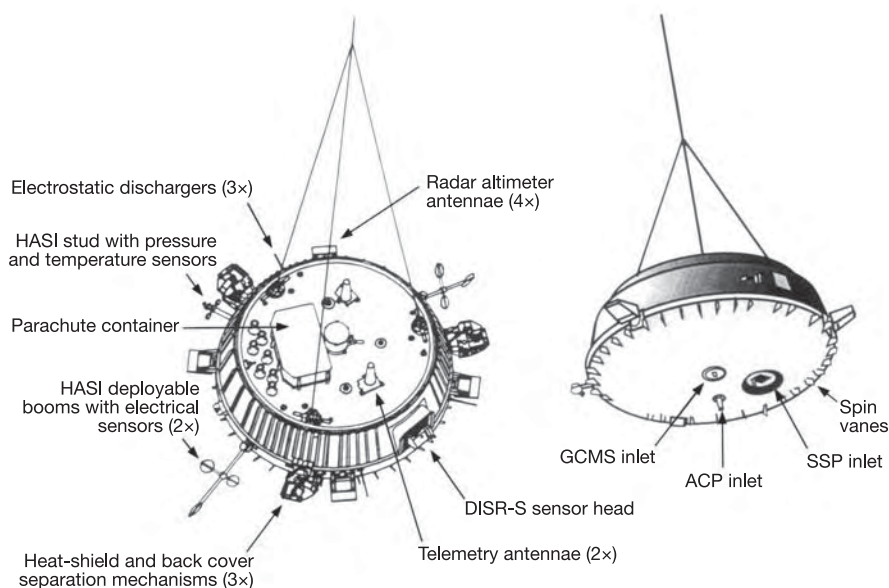


Figure 1 | Accommodation of the payload seen from two different perspectives. The external units of the probe are labelled. Five of the instruments required direct access to the gas flow and/or a clear field of view. The inlet ports of both the GCMS¹² and the ACP¹¹ were mounted close to the nose (apex) of the probe. The SSP¹³ also required direct access of the sensors in its 'Top Hat' structure to the gas—and eventually to a fluid if the Huygens probe had landed on a liquid surface. The judiciously designed SSP impact sensor protruded in front of the 'Top Hat' to allow direct detection of the impact a few milliseconds before the probe body itself reached the surface. The pressure and temperature sensors of the HASI⁹ were located on a fix stub and the electric properties sensors on two booms that were deployed immediately after the heat shield was released. The HASI microphone was

mounted on the outer ring. HASI included a sensitive accelerometer located near the centre of gravity of the probe in the entry configuration. The DISR¹⁴ sensor head was mounted on the outer rim of the probe ring so that it provided a clear field of view of almost 180 degrees from zenith to nadir. Probe spin allowed panoramic observations that took advantage of the probe rotation under the parachute. The Doppler Wind Experiment¹⁰ (DWE) included two ultra-stable oscillators designed to stabilize the transmitted carrier frequency and the corresponding receiver on one of the two radio links. The probe altitude was measured by a set of two radar altimeters that were switched on at an altitude of about 60 km, but started to provide useful measurements at an altitude of 45 km.

Composition measurements were made on the surface.⁴⁰Ar was detected on it. Its presence indicates that Titan has experienced in the past, and is probably still experiencing today, internal geologic activity. The time profile of the composition of surface vapours obtained by GCMS shows that the Huygens probe landed on a surface wet with methane, which evaporated as the cold soil was heated by the warmer probe. Compounds not seen in the atmosphere, such as C₆H₆, C₂N₂ and CO₂, were nevertheless detected in the gas from the surface material. Those measurements, which have not yet been fully analysed, appear to indicate complex chemical processes occurring on or in Titan's surface, as well as in the atmosphere.

The Huygens observations are presented and discussed in more detail in the accompanying papers^{15,18–22}. We now aim to put the Huygens mission operations into perspective.

The Huygens mission

Launch and flight to Saturn. The Cassini-Huygens spacecraft was launched from Cape Canaveral complex in Florida on 15 October 1997, with the probe mated onto the side of the orbiter. In this configuration, the orbiter provided electrical power to the probe through an umbilical connection. Commands and data were also exchanged by this route. During the seven-year journey to Saturn, the Huygens probe was subjected to 16 in-flight checkouts to monitor the health of its subsystems and scientific instruments¹⁶. During these in-flight tests, maintenance was performed and calibration data were obtained in preparation for the mission at Titan. The special in-flight tests designed to characterize the communication radio link

between the probe and the orbiter were especially important (see the Supplementary Information).

In the first link test in 2000, a flaw was discovered in the design of the Huygens telemetry receiver on board the orbiter that would have resulted in the loss of a large fraction of the Huygens probe's scientific data during the actual mission at Titan. Originally the Huygens mission was planned to be executed at the end of the first orbit around Saturn. As a remedy to the radio receiver flaw, the first two orbits of the original mission were redesigned^{7,8} into three shorter orbits that enabled the Huygens mission to be carried out on the third orbit. The re-designed orbiter trajectory during the probe relay is shown in Fig. 2. This trajectory provided a Doppler shift on the probe-orbiter radio link that was compatible with the well-characterized receiver performance and it also smoothly reconnected with the already-designed post-Huygens orbiter four-year trajectory⁸.

As a bonus, the new trajectory allowed early orbiter observations of Titan's upper atmosphere in order to validate the so-called Titan atmosphere engineering model⁷ well before the Huygens probe release. It led to improvements in our knowledge of the structure and the composition of the upper atmosphere; in particular, it provided better constraints on the argon concentration and indicated that methane was not present in sufficient quantity to affect the probe entry adversely (that is, via excessive radiative heating). Indeed, the new mission scenario led to the Huygens mission being completely successful. This achievement was the culmination of more than 20 years of work and shows that the in-flight rework of the mission was necessary and was successfully implemented.

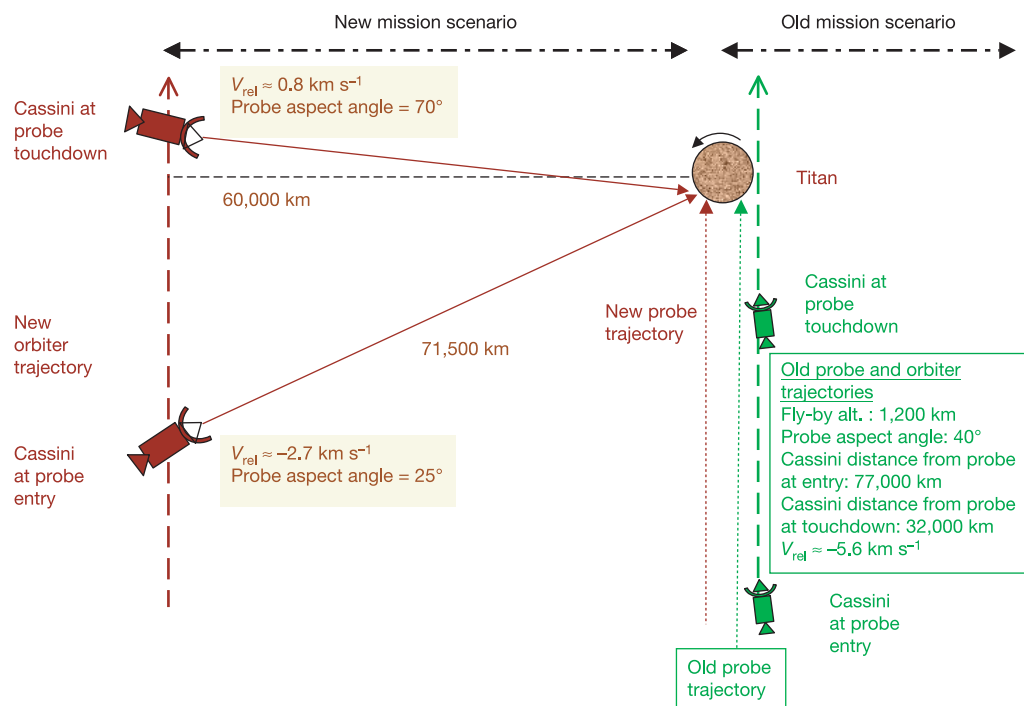


Figure 2 | Orbiter trajectory during the probe mission. This trajectory was implemented after a flaw was discovered in 2000 in the design of the Huygens telemetry receiver aboard Cassini. As originally designed, its telemetry demodulator was not able to receive and properly decode the transmissions at the expected frequency shift of about 25 p.p.m. (parts per million). The relative orbiter-probe velocity (Doppler shift) would have contributed 19 p.p.m., while the thermal frequency drift of the oscillator clocking the data stream would have contributed an additional 5–6 p.p.m. After full characterization of the receiver performance, a new mission scenario was designed to work around the constraints imposed by the receiver. The new design was developed in 2001 and was implemented during 2002–2004 (refs 7, 8, 16). The solution required a combination of the following measures: (1) A new Cassini trajectory that minimized the relative

probe-orbiter velocity. This changed the geometry of the Titan encounter by Cassini during the probe mission. It required the probe mission, initially planned to be conducted on the first orbit around Saturn, to be delayed until the third orbit. This trajectory change decreased the Doppler shift by 10–15 p.p.m. For reference, the old baseline trajectory is also indicated. (2) Pre-heating the Huygens probe before its arrival at Titan, by programming its wake-up four hours earlier than planned. As a result of the on-board oscillator that clocked the data stream frequency being warmer, the frequency of the data stream was further decreased (up to 3–4 p.p.m.). The pre-heating was implemented by appropriate changes in the on-board software of both the probe and the scientific instruments. It provided the robustness needed for the new mission.

Probe release. In preparation for releasing the probe, the Cassini-Huygens spacecraft had been set on a Titan-impact trajectory. Following its release, the Huygens probe had no manoeuvring capability and had to function autonomously. The Huygens release trajectory was achieved via a 'probe targeting manoeuvre' with a speed adjustment of 12 m s^{-1} on 17 December 2004, followed by a 'probe targeting clean-up manoeuvre' on 23 December 2004. After the separation of the Huygens probe on 25 December at 02:00 UTC, Cassini performed an 'orbiter deflection manoeuvre', so that it would not crash into Titan, and a 'clean-up manoeuvre' for final adjustment of its trajectory. These were on 28 December 2004 and 03 January 2005 respectively and placed Cassini on the correct trajectory for receiving data from the Huygens probe during the descent. The responsibilities for meeting the probe's trajectory requirements were shared between NASA/JPL and ESA. The targeting of the probe, the NASA/JPL responsibility, was specified at an altitude of 1,270 km, very close to the atmosphere's upper layer, above which no significant drag was expected. From this point onward ESA was responsible for the probe's trajectory.

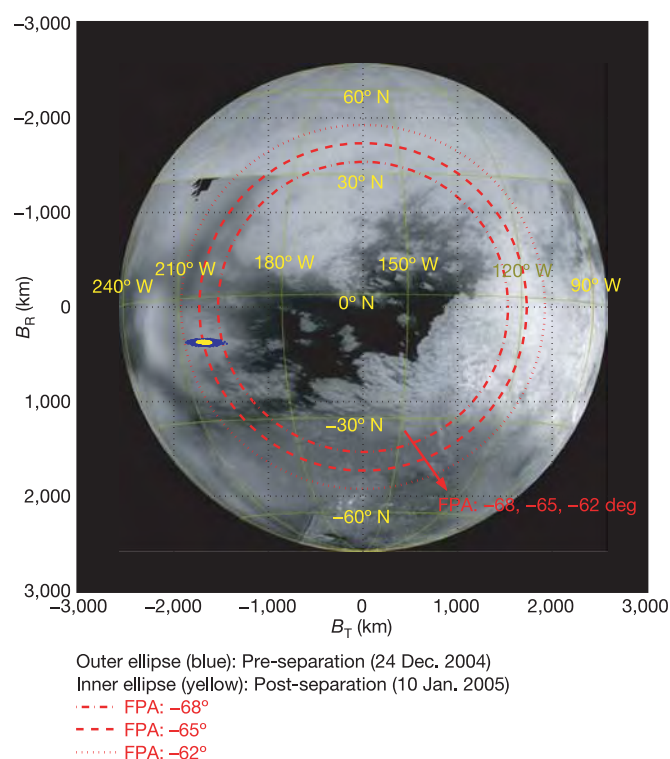


Figure 3 | Probe targeting as seen on a projection of the Titan disk. The surface image comes from the Cassini orbiter camera observation⁵ during the Titan fly-by on 26 October 2004. The three red curves give the targeted uncertainty ellipse of the entry point at an altitude of 1,270 km, for an entry angle of -62° (dotted line), -65° (solid line), and -68° (dashed line). (An entry angle of -90° would have given an entry point right in the centre of the figure). The blue ellipse gives the dispersion of the entry point as computed before the Cassini-Huygens separation, while the yellow ellipse indicates a reduced dispersion, as computed four days before the entry. The achieved probe Flight Path Angle (FPA) (the angle between the probe velocity vector and the local horizon) at the 1,270 km interface altitude was -65.4° with an uncertainty of $\pm 0.85^\circ$, compared to the requirement of -65° with an uncertainty of $\pm 3^\circ$ (99% confidence level). The projection is shown in the Titan B-plane (impact plane), which is defined as the plane perpendicular to the asymptotic approach velocity and passing through Titan's centre. The T-axis is contained in Titan's equatorial plane, and the R-axis is perpendicular to it. The time uncertainty related to the arrival at 1,270 km altitude was reconstructed during the post-flight analysis as 5.8 s. Background image adapted from JPL's photojournal image PIA06201. Courtesy NASA/JPL Space Science Institute.

The spring-loaded Huygens separation mechanism, called the Spin Eject Device, had three points of attachment to the probe. It provided a speed increment relative to the orbiter of 33 cm s^{-1} . The Spin Eject Device also imparted to the probe an anti-clockwise spin of 7.5 r.p.m. (when viewed from the orbiter). This provided inertial stability during the ballistic trajectory and atmospheric entry.

Coast and probe 'wake up'. The Huygens probe was set on a ballistic trajectory that took a little over 20 days. During this time, the probe was dormant, with only three redundant timers counting down to a specific time programmed to end 4 h and 23 min before the predicted entry. At this time, battery power was turned on and the on-board computers, their sensors (accelerometers, and later in the descent the radar altimeters), and the scientific instruments were energized according to the pre-programmed sequence. The probe 'woke up' as planned, at 04:41:33 UTC on 14 January 2005. The Huygens probe's receivers on board the Cassini orbiter were powered on from 06:50:45 to 13:37:32 UTC. The Huygens probe arrived at the 1,270 km interface altitude on the predicted trajectory (Fig. 3) on 14 January 2005 at 09:05:53 UTC, just a few seconds before the expected time.

Entry, descent and landing. The Huygens scientific mission proper took place during the entry, descent, landing and post-landing phases. Table 1 shows the list of the main mission events. The descent of the probe through Titan's atmosphere was controlled by parachutes. The aerodynamic conditions under which the main parachute had to be deployed were critical. The correct instant for parachute deployment (mission time event, t_0) (the nomenclature t_0 is equivalent to T_0 in some of the accompanying papers) was determined by the probe on-board computers that processed the measurements from the accelerometers that monitored the probe's deceleration¹⁶. Pyrotechnic devices fired a mortar that pulled out a pilot chute, which in turn removed the probe's back cover and pulled out the main parachute. Then, 30 s later, the front shield was released. It was expected that, by this time, the probe would have stabilized under the main parachute. During the entry phase, telemetry could not be transmitted by the probe until its back cover was removed. Thus, a limited set of engineering housekeeping data and the HASI science accelerometer data⁹ acquired during entry was stored on-board the probe for transmission to the orbiter after the radio link was established.

Post-flight data analysis showed that only one of the receivers (channel B) was phase-locked and functioned properly. Channel A had an anomaly that was later identified as being due to the

Table 1 | Huygens mission timeline on 14 January 2005

Activity	Time (h:min:s UTC)	Mission time, $t - t_0$ (h:min:s)
Probe power-on	04:41:18	-4:29:03
Probe support avionics power-on	06:50:45	-2:19:56
Arrival at interface altitude (1,270 km)	09:05:53	-0:04:28
t_0 (start of the descent sequence)	09:10:21	0:00:00
Main parachute deployment	09:10:23	0:00:02
Heat shield separation	09:10:53	0:00:32
Transmitter ON	09:11:06	0:00:45
GCMS inlet cap jettison	09:11:11	0:00:50
GCMS outlet cap jettison	09:11:19	0:00:58
HASI boom deployment (latest)	09:11:23	0:01:02
DISR cover jettison	09:11:27	0:01:06
ACP inlet cap jettison	09:12:51	0:01:30
Stabilizer parachute deployment	09:25:21	0:15:00
Radar altimeter power-on	09:42:17	0:31:56
DISR surface lamp on	11:36:06	2:25:45
Surface impact	11:38:11	2:27:50
End of Cassini-probe link	12:50:24	3:40:03
Probe support avionics power-off	13:37:32	4:27:11
Last channel A carrier signal reception	~14:53	5:42:39
by Earth-based radio telescopes	16:00 (ERT)	

The second column gives the time in UTC (for the probe), while the third column gives the time relative to t_0 , where t_0 is the official start of the descent associated with the pilot chute deployment event. ERT, Earth Received Time.

unfortunate omission of the telecommand to apply power to the ultra-stable oscillator driving the channel A receiver (see Box 1 for further details). Subsequent on-board events were determined by the on-board software that initiated a set of commands at times all related to the moment the pilot chute was released. These commands included switching on other instruments and the replacement of the main parachute by a smaller 'stabiliser chute' after 15 min, to ensure that the probe would reach the surface of Titan within the designed duration of the mission (150 min maximum for the descent under parachute).

The actual duration of the descent following the t_0 event was 2 h 27 min 50 s. During the first part of the descent, the probe followed the nominal time-based sequence with the instrument operations defined by commands in the on-board mission timeline. The later part of the descent sequence was optimized by taking into account the altitude measurements provided by two redundant radar altimeters^{7,16}. The altimeters were switched on 32 min after t_0 , which corresponded to an altitude of around 60 km. They provided altitude measurements to the on-board computers, which filtered and compared the measurements to the predicted altitude, in order to exclude erratic measurements at high altitude and to provide reliable measured altitude information to the payload instruments. This allowed for optimization of the measurements during the last part of the descent. The DISR measurements sequence was adjusted to measured altitude below 10 km and its lamp was switched on at 700 m above the surface¹⁹. The HASI and SSP instruments were set to their proximity and surface modes^{15,20} at low altitude above the surface.

The probe landed safely with a vertical speed of about 5 m s^{-1} and continued thereafter to transmit data for at least another 3 h 14 min, as determined by the detection and monitoring of the probe's 2.040-GHz carrier signal by the Earth-based radio telescopes. Throughout this time, Cassini was oriented to receive the two incoming radio signals from the probe by continuously pointing its high gain antenna to the predicted Huygens landing point. After listening for the longest possible duration of the Huygens probe's visibility, the orbiter was commanded to re-point its high gain antenna to Earth for transmission of the stored Huygens telemetry data. At that time, Cassini was at a distance of 1,207 million kilometres (8.07 AU) from the Earth (the one-way light-time was 67 min 6 s).

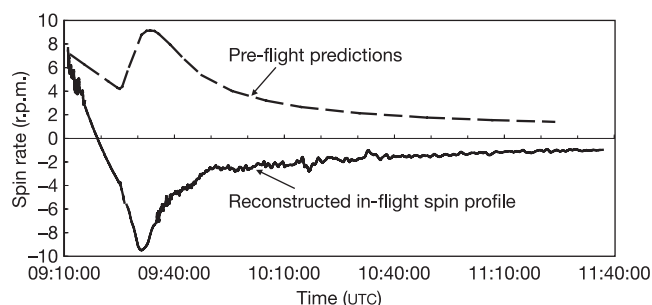


Figure 4 | Spin rate profile as a function of time. The solid curve displays the value derived from the radial accelerometer measurements and the spin phase variation of the automatic gain control of the probe-to-orbiter radio link. The probe entered the atmosphere and went through the entry with the expected spin rate (around 7.5 r.p.m.) in the anticlockwise direction. The spin rate decreased more rapidly than predicted under the main parachute and unexpectedly reversed direction after 10 min. It continued to spin with the expected rate but in the clockwise direction for the rest of the descent. The reason for this behaviour is under investigation. The post-flight verifications that could be made from design documentation do not show evidence for incorrect design or implementation of the spin vanes. Further detailed investigations of the aerodynamic interaction of the air flow with the probe under parachute may be required to explain this behaviour. The dashed line displays the predicted spin profile.

The data were received by the ground stations of the NASA Deep Space Network (DSN) and eventually delivered to the Huygens Probe Operations Centre (HPOC) in ESA's European Space Operation Centre (ESOC, Darmstadt, Germany) for science and engineering analysis. A 1-h margin was built into the orbiter sequence to cope with uncertainties as to when the orbiter would disappear below the horizon. As seen from the probe landing site, the orbiter actually set below the horizon at 12:50:24 UTC. The probe's channel A carrier signal was still being received on Earth by radio telescopes at the time of the planned completion of the observations, at 16:00 UTC (Earth received time), meaning that the probe was still operating at 14:53 UTC (Titan time). Post-flight analysis of the probe telemetry data indicates that the batteries probably became fully discharged at about 15:10 UTC, a mere 17 min after the Huygens radio signal was last verified on Earth. It is thought that the probe continued to function until the batteries were exhausted.

Trajectory reconstruction. The probe arrived at the 1,270 km interface altitude with the spin imparted at separation in the anti-clockwise direction. No significant spin modification was observed during the entry. The spin decreased more than expected under the main parachute and unexpectedly changed direction after 10 min. The probe continued spinning in the unexpected direction (clockwise) for the rest of the descent as illustrated in Fig. 4. No explanation was found for this behaviour, which is still under investigation.

Figure 5 shows the probe entry and descent altitude and vertical velocity profiles. The methodology that was used for the reconstruction effort is described in more detail in refs 24–27. The determination of the landing site coordinates is a complex and iterative task and requires several assumptions. At present, the best estimate, based on the combined Descent Trajectory Working Group (DTWG), DISR and DWE reconstruction, is a latitude of $10.3^\circ (\pm 0.4^\circ)$ south and a longitude of $167.7^\circ (\pm 0.5^\circ)$ east.

Summary and discussion. The probe and its scientific payload performed close to and sometimes beyond expectations. The in-flight

Box 1 | Channel A anomaly

The mission had two probe-orbiter radio link channels, which we refer to as channels A and B. Both transmitters (on board the probe) and both receivers (on board Cassini) were equipped with a temperature-controlled crystal oscillator (TCXO) which provided sufficient frequency stability ($\sim 10^{-6}$) for telemetry. One of the channels (channel A) was additionally equipped with ultra-stable oscillators (USOs) that were needed for the Doppler Wind Experiment (DWE)^{10,18}, which required a stable carrier frequency signal. As part of finalising the Huygens probe's configuration for its mission, it had been decided to use the channel A USOs instead of the TCXOs because the performance of the USOs had been very satisfactory during the seven-year cruise.

The command to power on the USO on the receiver side was unfortunately omitted. As a result, the Channel A receiver on board Cassini did not have a reference oscillator and was unable to lock onto the Huygens signal. Consequently, the frequency measurements for the Doppler Wind Experiment (DWE), together with the non-redundant telemetry data on Channel A, were lost.

The loss of the DWE data was, fortunately, largely mitigated by the radio astronomy segment of the mission consisting of a network of ground-based radio telescopes. The Channel A carrier signal, driven by the probe's USO, was received by 15 radio telescopes and tracked for post-flight data analysis. Real-time Doppler tracking information was obtained through the two largest telescopes of the network: the NRAO R. C. Byrd Green Bank Telescope (West Virginia, USA) and the CSIRO Parkes Radio Telescope (New South Wales, Australia). Both telescopes were equipped with NASA Deep Space Network's Radio Science Receivers (RSR) operated by the Radio Science Group of the Jet Propulsion Laboratory. In addition, the other 13 radio telescopes recorded the Channel A carrier signal for non-real-time Doppler and VLBI analysis.

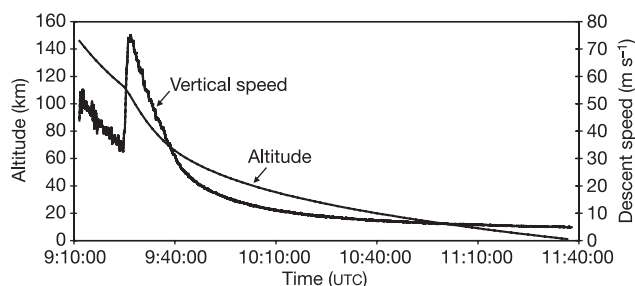


Figure 5 | Reconstructed altitude and descent speed as a function of mission time. The entry trajectory reconstruction from the official NASA/ESA interface point and epoch to the initiation of the parachute sequence t_0 is based on the numerical integration of the probe's equations of motion, which requires as an input the initial state vector (provided by the Cassini Navigation Team), the (modelled) gravitational force of the planet, the (measured) aerodynamic deceleration, and an assumption of a high-altitude wind speed profile (that is, 90 m s^{-1} prograde was assumed). The descent-phase reconstruction is based on the altitude and descent-speed reconstruction from the measured HASI temperature and pressure profiles¹⁵, the atmospheric mole fractions as measured by GCMS²¹, the impact time as most accurately measured by the SSP²⁰ internal accelerometer (ACC-I), and the DWE-derived¹⁸ zonal wind speed profile. Once the entry and descent trajectories were reconstructed independently from each other, a least-squares fitting algorithm was applied to adjust the probe initial conditions at the interface epoch and to ensure a smooth transition between the entry and descent trajectories.

modifications of the Huygens part of the mission, to cope with the receiver design flaw detected in 2000, was highly successful. The loss of data on channel A, due to a telecommand omission, was largely compensated for by the flawless transmission on channel B, with not a single bit missing until the radio link signal-to-noise decreased below the design limit of 3.3 dB, in the last 10 min of surface transmission, and the fact that the DWE scientific objectives were largely recovered by using data from the Earth-based radio telescope observations.

Deceleration and load levels measured during the hypersonic entry were well within the expected limits and all prime systems worked well, with no need to have recourse to the two back-up systems (g-switches) that had also been activated. The parachute performance was within the expected envelope, although the descent time, at slightly less than 2 h 28 min, was only just within the predicted envelope of 2 h 15 min $\pm$ 15 min. The descent was rather smooth under the main parachute but rougher than anticipated during the first hour under the last parachute. A detailed profile of the atmosphere is being worked out from the scientific measurements to allow the parachute performance to be studied in detail.

An exciting scientific data set was returned by the Huygens probe, offering a new view of Titan, which appears to have an extraordinarily Earth-like meteorology, geology and fluvial activity (in which methane would play the role of water on Earth). While many of Earth's familiar geophysical processes appear to occur on Titan, the chemistry involved is quite different. Instead of liquid water, Titan has liquid methane. Instead of silicate rocks, Titan has frozen water ice. Instead of dirt, Titan has hydrocarbon particles settling out of the atmosphere. Titan is an extraordinary world having Earth-like geophysical processes operating on exotic materials under very alien conditions²⁸. The Huygens data set provides the ground-truth reference for the interpretation of the remote observations of the Huygens landing site by orbiter instruments, and more generally the global observations of Titan. Future observations of the Huygens landing site by Cassini should allow us to place the local Huygens maps into their global context and are expected to tell us whether changes can be seen. Probe-orbiter synergistic studies are a key aspect for achieving the very ambitious Cassini-Huygens objectives at Titan.

Before the Huygens mission, it was thought that Titan could be a place of astro-biological interest^{29,30}. The Huygens results summarized in this paper and detailed in the papers that follow reveal the uniqueness of Titan in the Solar System as a planetary-scale laboratory for studying pre-biotic chemistry, which confirms the astro-biological interest of Saturn's largest moon. The exploration of Titan has just begun.

Received 19 June; accepted 21 October 2005.

Published online 30 November 2005.

- Smith, P. H. *et al.* Titan's surface revealed by HST imaging. *Icarus* **119**, 336–349 (1996).
- Meier, R., Smith, B. A., Owen, T. C. & Terrile, R. J. The surface of Titan from NICMOS observations with the Hubble Space Telescope. *Icarus* **145**, 462–473 (2000).
- Gibbard, S. G. *et al.* Titan: high-resolution speckle images from the Keck telescope. *Icarus* **139**, 189–201 (1999).
- Coustonis, A. *et al.* Maps of Titan's surface from 1 to 2.5 μm . *Icarus* **177**, 89–105 (2005).
- Porco, C. C. *et al.* Imaging of Titan from the Cassini spacecraft. *Nature* **434**, 156–165 (2005).
- Campbell, D. B., Black, G. J., Carter, L. M. & Ostro, S. J. Radar evidence for liquid surfaces on Titan. *Science* **302**, 431–434 (2003).
- Lebreton, J.-P. & Matson, D. L. The Huygens probe: science, payload and mission overview. *Space Sci. Rev.* **104**, 59–100 (2002).
- Matson, D. L., Spilker, L. J. & Lebreton, J.-P. The Cassini-Huygens mission to the saturnian system. *Space Sci. Rev.* **104**, 1–58 (2002).
- Fulchignoni, M. *et al.* The characterisation of Titan's atmospheric physical properties by the Huygens Atmospheric Structure Instrument (HASI). *Space Sci. Rev.* **104**, 395–431 (2002).
- Bird, M. K. *et al.* The Huygens Doppler Wind Experiment—Titan winds derived from probe radio frequency measurements. *Space Sci. Rev.* **104**, 613–640 (2002).
- Israel, G. *et al.* Huygens probe aerosol collector pyrolyser. *Space Sci. Rev.* **104**, 433–468 (2002).
- Niemann, H. B. *et al.* The gas chromatograph mass spectrometer for the Huygens probe. *Space Sci. Rev.* **104**, 553–591 (2002).
- Zarnecki, J. C. *et al.* Huygens' surface science package. *Space Sci. Rev.* **104**, 593–611 (2002).
- Tomasko, M. G. *et al.* The Descent Imager/Spectral Radiometer (DISR) experiment on the Huygens entry probe of Titan. *Space Sci. Rev.* **104**, 469–551 (2002).
- Fulchignoni, M. *et al.* *In situ* measurements of the physical characteristics of Titan's environment. *Nature* doi:10.1038/nature04314 (this issue).
- Clausen, K. C. *et al.* The Huygens probe system design. *Space Sci. Rev.* **104**, 155–189 (2002).
- Pogrebenko, S., *et al.* in *Proceedings of the International Workshop: Planetary Probe Atmospheric Entry and Descent Trajectory Analysis and Science* (6–9 October 2003, Lisbon) (ed. Wilson, A.) 197–204 (ESA SP-544, ESA Publications Division, Noordwijk, 2004).
- Bird, M. K. *et al.* The vertical profile of winds on Titan. *Nature* doi:10.1038/nature04060 (this issue).
- Tomasko, M. G. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* doi:10.1038/nature04126 (this issue).
- Zarnecki, J. C. *et al.* A soft solid surface on Titan as revealed by the Huygens Surface Science Package. *Nature* doi:10.1038/nature04211 (this issue).
- Niemann, H. B. *et al.* The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe. *Nature* doi:10.1038/nature04122 (this issue).
- Israel, G. *et al.* Complex organic matter in Titan's atmospheric aerosols from *in situ* pyrolysis and analysis. *Nature* doi:10.1038/nature04349 (this issue).
- Waite, H. *et al.* Ion neutral mass spectrometer results from the first flyby of Titan. *Science* **308**, 982–986 (2005).
- Atkinson, D. H., Kazeminejad, B., Gaborit, V., Ferri, F. & Lebreton, J.-P. Huygens probe entry and descent trajectory analysis and reconstruction techniques. *Planet. Space Sci.* **53**, 586–593 (2005).
- Kazeminejad, B. *Methodology Development for the Reconstruction of the ESA Huygens Probe Entry and Descent Trajectory*. PhD thesis, Karl-Franzens Univ. (2005).
- Kazeminejad, B. & Atkinson, D. H. The ESA Huygens probe entry and descent trajectory reconstruction. In *Proceedings of the International Workshop: Planetary Probe Atmospheric Entry and Descent Trajectory Analysis and Science* (6–9 October 2003, Lisbon) (ed. Wilson, A.) 137–149 (ESA SP-544, ESA Publications Division, Noordwijk, 2004).
- Kazeminejad, B. *et al.* Simulation and analysis of the revised Huygens Probe entry and descent trajectory and radio link model. *Planet. Space Sci.* **52**, 799–814 (2004).
- Lorenz, R. D. & Mitton, J. *Lifting Titan's Veil* (Cambridge Univ. Press, Cambridge, UK, 2002).
- Raulin, F. & Owen, T. Organic chemistry and exobiology on Titan. *Space Sci. Rev.* **104**, 377–394 (2002).

30. Schulze-Makuch, D. & Grinspoon, D. H. Biologically enhanced energy and carbon cycling on Titan. *Astrobiology* 5(4), 560–567 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The Cassini-Huygens mission is the result of an exemplary international collaboration in space exploration. Huygens involved more than 40 European industries and two US suppliers. The work of the members of the Cassini and Huygens teams from ESA, NASA/JPL, the Huygens industrial consortium led by Alcatel, and all Cassini-Huygens investigators is acknowledged. We especially acknowledge the Cassini orbiter teams that made their early observations available in advance to Huygens, and to R. Yelle for his leadership in coordinating the Titan Atmosphere Model Working Group. Special acknowledgements to B. Smeds for his work in designing and managing the Huygens link test that allowed the Doppler problem to be detected and solved,

to L. Popken for his modelling of the Huygens Digital Radio Receiver and to the whole Huygens recovery task force led by K. Clausen and L. Deutsch. We thank K. van't Klooster for his efforts to initiate and promote the Huygens VLBI experiment, and J. Louet for his support. The Earth-Based Huygens Doppler tracking experiment is led by W. Folkner. We appreciated the support provided by the National Radio Astronomy Observatory (NRAO), operated by Associated Universities Inc., under a cooperative agreement with the NSF, and the one provided by the Australia Telescope National Facility (ATNF) managed by CSIRO. We also thank M. Bird, R. Lorenz, R. A. Preston and J. C. Zarnecki for a careful reading of various versions of the manuscript and for providing comments.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.-P.L. (jean-pierre.lebreton@esa.int).

Rain, winds and haze during the Huygens probe's descent to Titan's surface

M. G. Tomasko¹, B. Archinal², T. Becker², B. Bézard³, M. Bushroee¹, M. Combes³, D. Cook², A. Coustenis³, C. de Bergh³, L. E. Dafoe¹, L. Doose¹, S. Douté⁴, A. Eibl¹, S. Engel¹, F. Gliem⁵, B. Grieger⁶, K. Holso¹, E. Howington-Kraus², E. Karkoschka¹, H. U. Keller⁶, R. Kirk², R. Kramm⁶, M. Küppers⁶, P. Lanagan¹, E. Lellouch³, M. Lemmon⁷, J. Lunine^{1,8}, E. McFarlane¹, J. Moores¹, G. M. Prout¹, B. Rizk¹, M. Rosiek², P. Rueffer⁵, S. E. Schröder⁶, B. Schmitt⁴, C. See¹, P. Smith¹, L. Soderblom², N. Thomas⁹ & R. West¹⁰

The irreversible conversion of methane into higher hydrocarbons in Titan's stratosphere implies a surface or subsurface methane reservoir. Recent measurements from the cameras aboard the Cassini orbiter fail to see a global reservoir, but the methane and smog in Titan's atmosphere impedes the search for hydrocarbons on the surface. Here we report spectra and high-resolution images obtained by the Huygens Probe Descent Imager/Spectral Radiometer instrument in Titan's atmosphere. Although these images do not show liquid hydrocarbon pools on the surface, they do reveal the traces of once flowing liquid. Surprisingly like Earth, the brighter highland regions show complex systems draining into flat, dark lowlands. Images taken after landing are of a dry riverbed. The infrared reflectance spectrum measured for the surface is unlike any other in the Solar System; there is a red slope in the optical range that is consistent with an organic material such as tholins, and absorption from water ice is seen. However, a blue slope in the near-infrared suggests another, unknown constituent. The number density of haze particles increases by a factor of just a few from an altitude of 150 km to the surface, with no clear space below the tropopause. The methane relative humidity near the surface is 50 per cent.

The surface of Titan has long been studied with various instruments, including those on the Hubble Space Telescope (HST) and ground-based adaptive optics systems¹. More recently, Cassini investigations using the charge-coupled device (CCD) camera², the Visible and Infrared Mapping Spectrometer (VIMS) instrument³, and the Radio Detection and Ranging (RADAR) imaging system⁴ have provided more detailed views of Titan's surface in the hope of revealing how the methane in Titan's atmosphere is replenished from the surface or interior of Titan. Of the Cassini imagers, the Imaging Science Subsystem (ISS) camera is potentially capable of the greatest spatial resolution, but Titan's obscuring haze limits its resolution on the surface to about 1 km, a value roughly similar to that available from VIMS and the radar imaging system. At this resolution, the bright and dark regions observed on the surface of Titan have proved difficult to interpret. Owing to its proximity to the surface, the Descent Imager/Spectral Radiometer (DISR) camera on the Huygens probe was capable of a linear resolution of some metres from a height of 10 km. In addition, the lower the probe descended, the less haze lay between the camera and the ground. The DISR was capable of linear resolution orders of magnitude better than has been available from orbit, although of a much smaller portion of Titan's surface. Also, a lamp was used at low altitude to measure the continuous reflectance spectrum of the surface without the complications introduced by observations through large amounts of methane and aerosol haze⁵.

In addition to studying the surface of Titan, the DISR took

measurements of solar radiation in the atmosphere. Spectrometers looking upward at continuum wavelengths (between the major methane absorptions) as well as downward measured the vertical distribution and wavelength dependence of the aerosol haze opacity. Measurements of the polarization of light at a scattering angle of 90° constrained the small dimension of the haze particles. Measurements of the brightness in the solar aureole around the Sun determined the projected area of the haze particles. Observations in the methane bands determined the methane mole fraction profile.

Data collection during the descent proceeded mostly, although not exactly, as planned. Turbulence during the first half of the descent tipped the probe more rapidly than expected, causing the Sun sensor to remain locked on the azimuth of the Sun for only a few successive rotations at a time. Below about 35 km, the signal from the direct solar beam was lost by the Sun sensor owing to the unexpectedly low temperature of this detector. These effects caused data from each of the DISR sub-instruments to be collected at mostly random, instead of specific, azimuths. Additionally, the probe rotated in the intended direction for only the first ten minutes before rotating in the opposite sense for the remainder of the descent. This resulted in ineffective baffling of the direct solar beam for the upward-looking visible spectrometer and the solar aureole camera. Consequently, some measurements made by the solar aureole camera are saturated, and the separation of the direct and diffuse solar beams in the visible spectral measurements must be postponed until a good model of the probe attitude versus time is available. Finally, the loss of one of

¹Lunar and Planetary Laboratory, University of Arizona, 1629 E. University Blvd, Tucson, Arizona 85721-0092, USA. ²US Geological Survey, Astrogeology, 2225 N. Gemini Drive, Flagstaff, Arizona 86001, USA. ³LESIA, Observatoire de Paris, 5 place Janssen, 92195 Meudon, France. ⁴Laboratoire de Planétologie de Grenoble, CNRS-UJF, BP 53, 38041 Grenoble, France. ⁵Technical University of Braunschweig, Hans-Sommer-Str. 66, D-38106 Braunschweig, Germany. ⁶Max Planck Institute for Solar System Research, Max-Planck-Str. 2, D-37191 Katlenburg-Lindau, Germany. ⁷Department of Physics, Texas A&M University, College Station, Texas 77843-3150, USA. ⁸Istituto Nazionale di Astrofisica — Istituto di Fisica dello Spazio Interplanetario (INAF-IFSI ARTOV), Via del Cavaliere, 100, 00133 Roma, Italia. ⁹Department of Physics, University of Bern, Sidlerstr. 5, CH-3012 Bern, Switzerland. ¹⁰Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA.

the two radio communication channels in the probe receiver aboard the orbiter resulted in the loss of half the images as well as several other low-altitude spectrometer measurements.

Despite these misfortunes, the DISR instrument collected a unique and very useful data set. Images of the surface with unprecedented resolution were collected over the boundary between bright and dark terrain seen from the orbiter. Owing to redundant transmission over both communication channels during most of the descent, almost all of the spectral and solar aureole observations were received. A very large set of high-quality spectra were obtained with good altitude resolution and with good coverage in azimuth both away from and towards the Sun. The images, the spectra, the Sun sensor pulses, the recording of the gain in the Cassini radio receiver, and information from Very Long Baseline Interferometry (VLBI) observations from Earth together will permit reconstruction of the probe attitude relative to the Sun as a function of time during the descent, enabling

a full analysis of the spectral data. The large number of solar aureole measurements included several acquired near the Sun and many polarization measurements opposite to the Sun. The surface science lamp worked exactly as planned, permitting surface reflection measurements even in strong methane absorption bands. Operations after landing included the collection of successive images as well as spectral reflectance measurements of the surface illuminated by the lamp from an assumed height of roughly 30 cm.

Taken together, the new observations shed substantial light on the role played by methane in forming the surface of Titan and how it is recycled into the atmosphere. The substantial relative humidity of methane at the surface and the obvious evidence of fluid flow on the surface provide evidence for precipitation of methane onto the surface and subsequent evaporation. Some indications of cryovolcanic flows are also seen. The vertical distribution and optical properties of Titan's haze have been characterized to aid the interpretation of remote measurements of the spectral reflection of the surface. The speed and direction of Titan's winds has also been measured for comparison with future dynamical models that include the radiative heating and cooling rates implied by the haze.

Physical processes that form the surface

The imagers provided views of Titan's previously unseen surface, thus allowing a deeper understanding of the moon's geology. The three DISR cameras were designed to provide overlapping coverage for an unbroken 360°-wide swath stretching from nadir angles between 6° and 96°. Some 20 sets of such images were planned during the descent. Because of the opacity of the haze in the passband of our imager, surface features could be discerned in the images only below about 50 km, limiting the number of independent panoramic mosaics that can be made of the surface. The loss of half of the images meant that Titan's surface was not covered by systematic overlapping triplets, as expected. Three different views of Titan's surface are

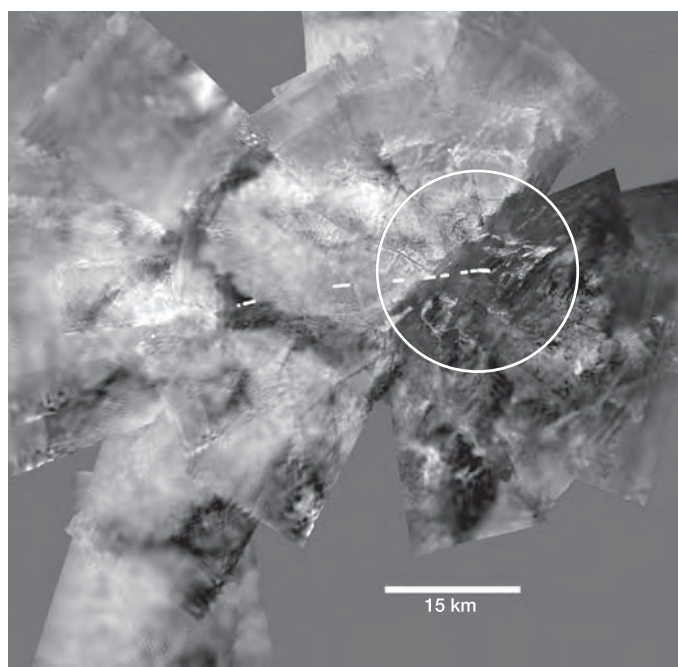


Figure 1 | View of Titan from 34 km above its surface. High-altitude (49 to 20 km) panoramic mosaic constructed from the DISR High and Medium Resolution Imagers (HRI and MRI) as projected from 34 km. The preliminary ground-track solution (indicated as small white points on gnomonic ground projection) represents the location of the probe when data were collected; north points up; scale indicated (although subsequent analysis indicates that north lies some 5–10° to the left of straight up in this and the two subsequent figures). Starting from the first surface image at 49 km, the probe moves in an east-northeastwardly direction at an initial speed of 20 m s⁻¹. Brighter regions separated by lanes or lineaments of darker material are seen. No obvious crater-like features are visible. The circle indicates the outline of the next-lowest pan, in Fig. 2. The method used for construction of panoramic mosaics incorporates knowledge of the probe's spatial location (longitude, latitude and altitude) and attitude (roll, pitch and yaw) at each image. With the exception of altitude, provided by the Huygens Atmospheric Structure Instrument¹² pressure sensor, none of these variables was directly measured. They are found through an iterative process in which a panorama is created, providing an improved ground-track and azimuth model, which results in an upgraded trajectory, which can improve the panorama, and so on. The current lack of pitch and roll knowledge constitutes the main source of error in the current composition and quality of the panoramas as well as the ground-track and wind speed determination reported below. Vigorous contrast-stretching in the images is required to reveal details washed out by the haze particle density at all altitudes in Titan's atmosphere. This contrast-stretching also displays the occasional ringing of the Discrete Cosine Transform data compressor, which appear as regular lines of bright and dark patterns, particularly in the MRI images.

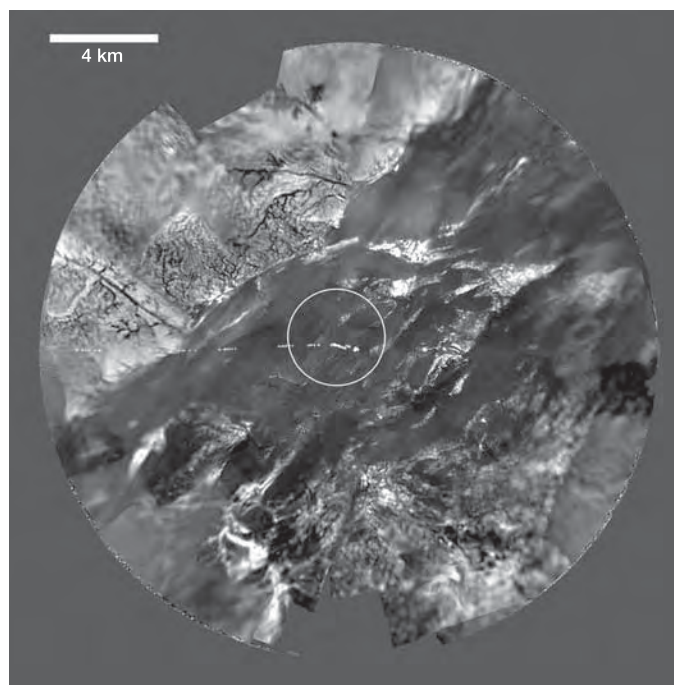


Figure 2 | View from 8 km. Medium-altitude (17 to 8 km) panoramic mosaic projected from 8 km. As in Fig. 1, the preliminary ground-track solution is indicated as points; north is up; scale indicated. At 11 km, the wind direction is at 0° (eastward), reaching -20° (southeastward) at an altitude of 8.5 km. The narrow dark lineaments, interpreted as channels, cut brighter terrain. The circle indicates the outline of the low-altitude pan in Fig. 3.

shown in Figs 1 to 3. A view of Titan's surface after the Huygens landing is shown in Fig. 4.

The highest view (Fig. 1), projected from an altitude of 34 km, displays an albedo variation very similar to the highest-resolution images provided by the ISS or VIMS cameras on the Cassini orbiter. It shows Titan's surface to consist of brighter regions separated by lanes or lineaments of darker material. No obvious impact features are visible, although craters less than roughly 10 km should not be abundant as a result of atmospheric shielding⁶. In the rightmost (eastern) part of the mosaic the images become sharper as the lower altitude of the camera causes the scale to decrease and the contrast to increase. More than a dozen brighter areas in that region seem to be elongated along a direction parallel with the main bright/dark boundary of that region. At the limit of resolution, narrow dark channels cut the bright terrain.

The next view, Fig. 2, projected from an altitude of 8 km, reveals a large number of these channels (detailed in Fig. 5). The channel networks have two distinct patterns: short stubby features and dendritic features with many branches. The region of the stubby network towards the west is significantly brighter than the dendritic region. The stubby channels are shorter, wider and relatively straight. They are associated with and often begin or end in dark circular areas, which suggest ponds or perhaps pits. The morphology of rectilinear networks with stubby heads is consistent with spring-fed channels or arroyos.

The dendritic network is consistent with rainfall drainage channels, implying a distributed rather than a localized source of a low-viscosity liquid. Stereo analysis of the dendritic region indicates an elevation of 50–200 m relative to the large darker plain to the south. It suggests that the brighter areas within the darker terrain are higher as well. The topographic differences are evident in Figs 6

and 7, which are three-dimensional renderings of the area just north of the landing site produced from the DISR images. They include the major bright–dark interface seen from above in Fig. 5. Figure 2 depicts many examples of these darker lanes of material between topographically higher, brighter areas. In fact the low contrast of the lowland plane argues that the entire dark region floods, and as the liquid drains the local topography drives flows as seen in the images. If the darker region is interpreted as a dried lakebed, it is too large to have been caused by the creeks and channels visible in the images. It may have been created by other larger river systems or some large-scale catastrophic event, which predates deposition by the rivers seen in these images.

The interpretation of the dark lanes within the brighter highlands as drainage features is so compelling as to dominate subsequent interpretation of other areas of images such as Figs 2 and 3. The prevailing bright–dark boundary of the region becomes a coastline, the bright areas separated from this boundary become islands. Bright streaks running parallel to the albedo boundary may be drift deposits or spalls fractured off the bright highlands owing to faulting along the shoreline.

When coupled with Fig. 4, which is an image of a typical offshore dark region, it is clear that the analogy has a limit. At present there is no liquid in the large dark lakebed imaged in Figs 1 to 5. The bright lobate feature, split by an apparently straight dark lane in the western part of the mosaic in Fig. 2, is a possible fissure-fed cryovolcanic flow. However, Fig. 4 also reveals rocks which—whether made of silicates or, more probably hydrocarbon-coated water-ice—appear to be rounded, size-selected and size-layered as though located in the bed of a stream within the large dark lakebed. Rounded stones approximately 15 cm in diameter and probably composed of water-ice, lie on top of a darker, finer-grained surface.

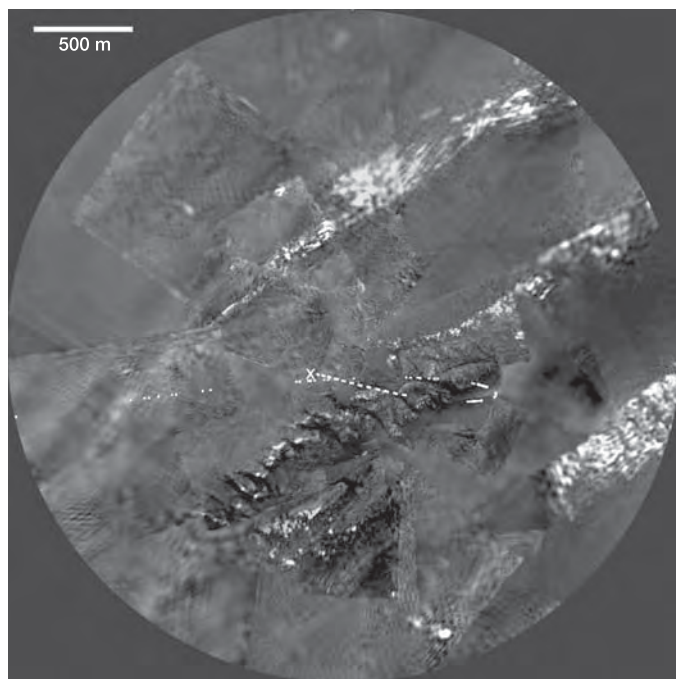


Figure 3 | View from 1.2 km. Low-altitude (7 to 0.5 km) panoramic mosaic projected from 1,200 m. As in Figs 1 and 2, the preliminary ground track is indicated as points; north is up; scale indicated. The probe's steady east-northeast drift halts altogether at an altitude of 7 km and reverses, moving west-northwest for some 1 km during the last 15 min of descent. Note the ridge near the centre, cut by a dozen darker lanes or channels. The projected landing site is marked with an 'X' near the continuation of one of the channels, whose direction matches the orientation of the stream-like clearing in the near-foreground of the southward-looking surface image, Fig. 4.

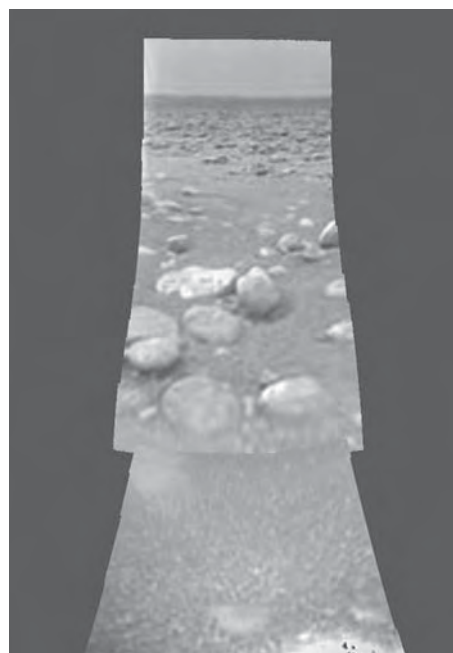


Figure 4 | The view from Titan's surface. Merged MRI and SLI images acquired after the Huygens probe soft-landing. Horizon position implies a pitch of the DISR nose upward by $1.7 \pm 0.2^\circ$ with no measurable roll. 'Stones' 10–15 cm in size lie above darker, finer-grained substrate in a variable spatial distribution. Brightening of the upper left side of several rocks suggests solar illumination from that direction, implying a southerly view, which agrees with preliminary evidence from other data sets. A region with a relatively low number of rocks lies between clusters of rocks in the foreground and the background and matches the general orientation of channel-like features in the low-altitude pan of Fig. 3. The bright spot in the lower right corner is the illumination of the DISR surface science lamp.

It is interesting to compare the brightness and colour of the scene shown in Fig. 4 with that of a similar scene on the Earth. The brightness of the surface of the Earth illuminated by full sunlight is about half a million times greater than when illuminated by a full moon. The brightness of the surface of Titan is about a thousand times dimmer than full solar illumination on the Earth (or 500 times brighter than illumination by full moonlight). That is, the illumination level is about that experienced about 10 min after sunset on the Earth. The colour of the sky and the scene on Titan is rather orange due to the much greater attenuation of blue light by Titan's haze relative to red light. If the Sun is high in the sky, it is visible as a small, bright spot, ten times smaller than the solar disk seen from Earth, comparable in size and brightness to a car headlight seen from about 150 m away. The Sun casts sharp shadows, but of low contrast, because some 90% of the illumination comes from the sky. If the Sun is low in the sky, it is not visible.

The sizes of the more than 50 stones in the image in Fig. 4 vary between 3 mm in diameter, the resolution limit of the imager, and 15 cm. No rocks larger than 15 cm are seen. The resolution of the last images taken before landing from a height of 200–300 m would be sufficient to identify metre-sized objects, and none are seen in the 40×35 m field of view. Figure 8 shows the R value, a measure of the fraction of the surface covered by rocks of a given size frequently used to describe the size distribution of impact craters or crater ejecta. A larger fraction of the surface is covered with rocks greater than 5 cm as opposed to smaller pebbles. The dominance of the cobbles 5–15 cm in size suggests that rocks larger than ~ 15 cm cannot be transported to the lakebed, while small pebbles (< 5 cm) are quickly removed from the surface. Figure 8 confirms the visual impression given by Fig. 4 that the surface coverage of rocks in the foreground of the image (< 80 cm horizontal distance from the probe) is higher than in the region beyond (about 80–160 cm). However, this trend is not seen for the pebbles less than 5 cm in size.

Elongated dark trails aligned with the general trend of the possible stream-bed visible in the centre of Fig. 4 extend from several of the distant boulders. The direction of the trails agrees with the general northwest–southeast alignment of the stream-like features shown in Fig. 3, because the last upward-looking spectra indicate that the probe settled with DISR facing southward. Images taken from the surface show no traces of the landing of the probe. The viewing

direction is probably generally not downwind (the parachute is not visible).

When coupled with the shapes, size selection and layering of the stones in Fig. 4, the elongated islands and their orientation parallel to the coastline in Fig. 1, the stubby and dendritic channel networks, as well as the ponds in Fig. 2 and Fig. 5, the major elements of the Titan surface albedo variations can be interpreted to be controlled by flow of low-viscosity fluids driven by topographic variation, whether caused by precipitation (the dendritic networks) or spring-fed flows (the stubby networks). We thus interpret the bright–dark albedo difference as follows: irrigation of the bright terrain results in darker material being removed and carried into the channels, which discharge it into the region offshore, thereby darkening it. Eolian processes such as wind gusts coupled with Titan's low gravity (compared to Earth) may aid this migration.

The dark channels visible in the lowest panorama (Fig. 3) seem to suggest south-easterly fluid flow across the lower plane, depositing or exposing the brighter materials (water ice?) along the upstream faces of the ridges.

Stereographic rendering of the dendritic channels just north of the probe landing site (Figs 6 and 7) shows that the slopes in bright terrain being dissected by the putative methane river channels are

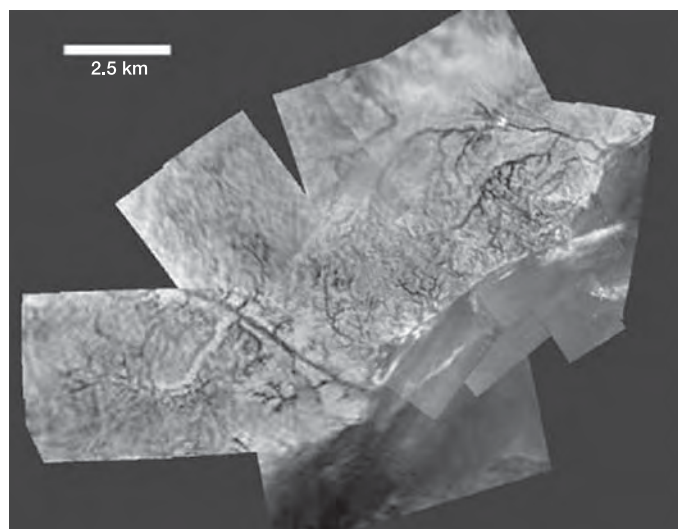


Figure 5 | View of 'shoreline' and channels. Panoramic mosaic projected from 6.5 km, showing expanded view of the highlands and bright–dark interface. As in previous figures, north is up; scale indicated. Branching and rectilinear channel networks of dark lanes are shown along an albedo boundary approximately 12 km long.

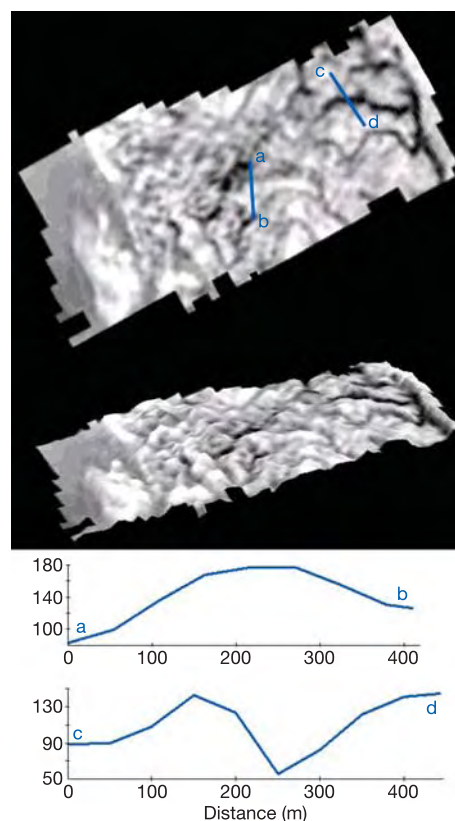


Figure 6 | Topographic model of highland region ~ 5 km north of the Huygens landing site. The top panel shows an orthorectified HRI image from stereo pair (vertical view). The middle panel shows a perspective view of the topographic model with $\sim 50^\circ$ tilt angle. No vertical exaggeration was applied (it is 1:1). The bottom panel shows profiles (a–b and c–d from the top panel) that illustrate the extremely steep topography in the region dissected by the drainages. All dimensions are in metres. A DISR stereo pair (using HRI frame 450 and MRI frame 601) was photogrammetrically analysed using a digital stereo workstation. The overlapping area of stereo coverage is about 1×3 km; the convergence angle is $\sim 25^\circ$. The coincidence of the drainage patterns with the valley floors gives confidence in the reality of the topographic model; the height accuracy is ~ 20 m. This preliminary model has been arbitrarily levelled so that the elevation differences are only relative.

extremely rugged; slopes of the order of 30° are common. This suggests relatively rapid erosion by flows in the river beds, resulting in the deeply incised valleys. Erosion by steep landslides on slopes approaching the angle of repose is probably the primary mechanism by which the rugged topography is formed. Figure 7 shows two stereographic views of the shoreline and hillside north of the landing site.

The wind profile

Assembly of the panoramic mosaics leads to the construction of a descent trajectory as part of an iterative process of image reconstruction. The trajectory can be used to derive the probe ground track and extract the implied wind velocity as a function of altitude. Correlation of the roughly 200 usable images acquired by DISR during its descent resulted in longitude and latitude values versus time, displayed in Fig. 9. Fitted by polynomials, these ground tracks were differentiated with respect to time and scaled to derive the horizontal wind speed and direction as functions of altitude.

The results indicate that the probe's steadily increasing eastward drift caused by Titan's prograde winds, as shown in Fig. 10, slowed from near 30 to 10 m s^{-1} between altitudes of 50 and 30 km and slowed more rapidly (from 10 to 4 m s^{-1}) between altitudes of 30 and 20 km . The winds drop to zero and reverse at around 7 km , near the expected top of the planetary boundary layer, producing a west-northwestwardly motion extending for about 1 km during the last 15 min of the descent (see Fig. 3). The generally prograde nature of the winds between 50 and 10 km agrees with models of Titan's zonal winds available before the arrival of Cassini or the Huygens probe⁷, although the wind speed is somewhat less than the average predicted before entry.

The planetary boundary layer is calculated to have a thickness of between 4 and 8 km , based on scaling the Earth's near-equatorial planetary-boundary-layer thickness of $1\text{--}2 \text{ km}$ by the inverse square-rooted ratio of the planetary rotation rates. The minimum horizontal wind speed at 7 km can thus be an indication of entry into the boundary layer⁸. The reversal of wind direction at this altitude is also consistent with the Voyager-derived equatorial temperature profile⁹, wherein the temperature gradient changes from dry adiabatic to a sub-adiabatic temperature gradient above 4 km altitude, indicating the top of the boundary layer.

The current ground-track and wind-speed analysis predicts

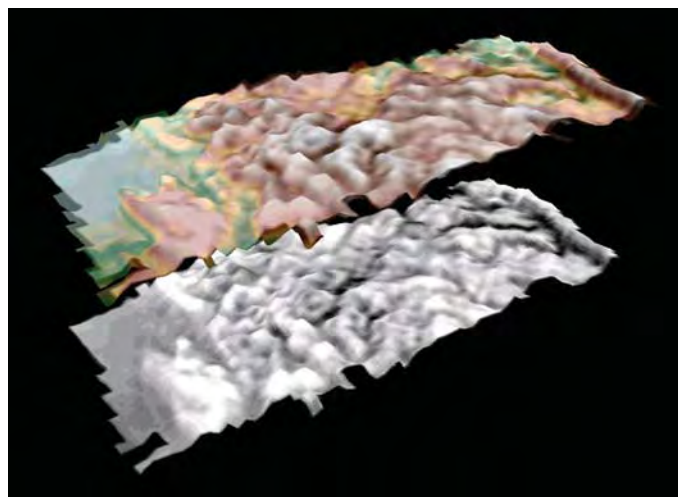


Figure 7 | Titan's surface. Perspective view of Titan's surface using a topographic model of the highland region $\sim 5 \text{ km}$ north of the Huygens probe landing site derived from the DISR images. The model in greyscale and false colour shows the elevation (pale white highest). The lowland plane or lakebed is to the left side of the display (in blue); the northern highlands (with the dendritic channels) is to the right.

winds of about 0.3 to 1 m s^{-1} near the surface. This velocity can be produced by any number of sources including pressure and temperature gradients and tides¹⁰.

Migration of surface material

The acquisition of visible spectra at known locations in the images allowed correlation of the reflectance spectra with different types of terrain. The Downward-Looking Visible Spectrometer (DLVS) was an imaging spectrometer measuring light between 480 and 960 nm as it projected the image of the slit onto the ground into up to 20 spatial resolution elements for nadir angles from 10° to 50° . Spectra were collected at nominally the same azimuths as the images, though often at slightly different altitudes (on different probe rotations). Interpolation between the times at which the spectral and image data were obtained located the spectra within the images. Determination of the surface reflectivity was hindered by scattering from the haze between the camera and the surface as well as by methane absorption. Correlation of the spectra with images was therefore best performed on measurements during which the altitude changed only slightly.

The centre of the image in Fig. 11 is displayed in true colour (that is, as the human eye would see it under Titan's atmosphere) using actual spectral data from one panorama. The area between the spectra is interpolated in azimuth. The coverage with spectra is similar to that shown in Fig. 12. The orange colour is due mainly to the illumination of the surface. Scattering and absorption (which dominate in the blue) cause the perceived true colour of the surface to change from yellow to orange with decreasing altitude. Note that the passband of the cameras peaked in the near infrared (at 750 nm), and therefore the brightness variations in the images would not necessarily be seen by the human eye.

In Fig. 12 the images are correlated with the ratio of the intensity in two methane windows ($827 \text{ nm}/751 \text{ nm}$) located in the infrared part of the spectrum where scattering is minimal and the systematic variability with nadir angle can be ignored. Reddening (high $827 \text{ nm}/751 \text{ nm}$ ratio) is concentrated at the area covered with drainage

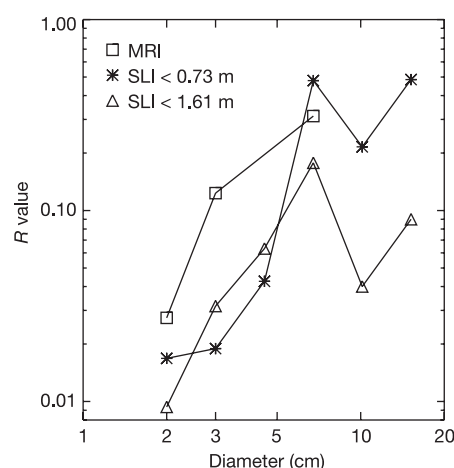


Figure 8 | Distribution of rock on the surface. Rocks larger than 1.63 cm as an R -plot, frequently used to describe size distribution of impact craters or crater ejecta. If N is the number of rocks per centimetre increment of rock size, the fraction of the surface area A covered by rocks with diameters between d and $d + \Delta d$ is approximately $N \times \Delta d \times d^2/A$. By keeping the size bin Δd proportional to the diameter d , the quantity $N \times d^3/A$ (the R value) is also proportional to the surface fraction covered by rocks of diameter d (with a proportionality constant of ~ 3). The plot shows R values derived from rock counts from the SLI and MRI surface images. For the SLI, R values from counts up to a distance from the probe of 73 cm and up to 161 cm are presented in separate curves. The comparison between the two curves suggests that the count is complete in the displayed size range. The increase of the R value with size corresponds to a higher fraction of the surface covered with large rocks than with smaller ones.

channels (north and northwest in the pan of Fig. 12) and to a lesser degree to the lake area adjacent to the coastline. The lake area in the southeast is not reddened. A preliminary analysis of spectra recorded in other panoramic cycles indicates that the land area in the north-east, which is not covered by drainage channels, is only moderately reddened compared to the river area. The reddening is not restricted to these two methane windows. Figure 13 shows that, in fact, it is present over the whole visual range, amounting to about 6% per 100 nm (note that atmospheric backscatter dominates over surface reflection at wavelengths below 600 nm).

The DLVS data clearly show that the highlands (high-albedo area) are redder than the lakebed (low-albedo area). Spectra of the lakebed just south of the coastline are less red than the highlands but clearly more red than the lakebed further away (that is, to the southeast). The data suggests that the brighter (redder) material of the hilly area may be of local origin, and is corrugated by rivers and drainage channels, and that the darker material (less red) is a substance that seems to be washed from the hills into the lakebed. It could be connected to the alteration of the highland terrain, either by precipitation, wind and/or cryoactivity. Additionally, it could indicate that the surface of the lowland area may be covered by different materials in regions that exhibit diverse morphology.

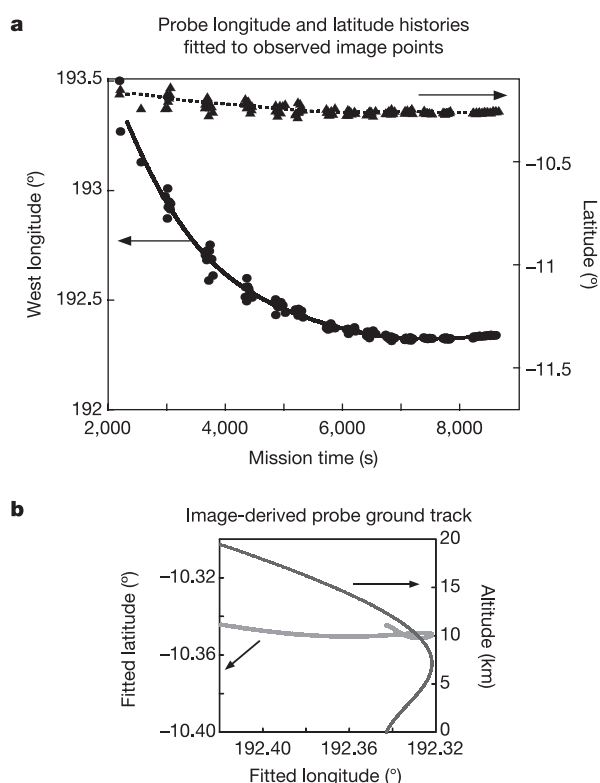


Figure 9 | Probe ground track. **a**, Sub-probe west longitude and latitude histories of the Huygens probe derived from panoramic image reconstructions. Arrows indicate the appropriate vertical axis. The image data points from which the latitude and longitude were derived are shown as triangles and dots respectively. The dotted and solid lines show polynomial fits to the data. Results adjusted to agree with the Descent Trajectory Working Group (DTWG-3) values at 2,200 s after T_0 (mission time). **b**, Probe longitude versus latitude (thicker line) and versus altitude (thinner line). The altitude axis needs to be expanded by a factor of almost six to recover one-to-one correspondence on a linear scale because the total longitudinal variation is less than 4 km. Using the Doppler Wind Experiment's (DWE) high-altitude references³³, the touchdown point (the predicted landing site) was extrapolated to west longitude 192.34°, latitude -10.34°. Using DTWG-3 high-altitude references³³, the touchdown point was extrapolated to 192.36°, latitude -10.36°.

Surface reflectivity and methane mole fraction

Spectra taken near Titan's surface allow measurement of its reflectance and determination of the local methane mole fraction. These measurements provide clues as to the make-up of Titan's crust. The Downward-Looking and Upward-Looking Infrared Spectrometers (DLIS and ULIS) cover the region from 840 to 1,700 nm with a resolution of 15 to 20 nm. The ULIS looks up at a half-hemisphere through a diffuser while the DLIS projects its slit into a 3 by 9° field centred at a 20° nadir angle. Below 700 m altitude, a 20-W lamp was turned on to illuminate the surface at wavelengths where solar light had been completely absorbed by methane in Titan's atmosphere. At low altitudes we took repeated DLIS and ULIS spectra at short integration times (1 s). Nine DLIS and seven ULIS spectra were received between 734 and 21 m altitude. The DLIS continued to measure surface spectra free of atmospheric methane absorption after landing. About 20 such identical spectra were acquired from a distance of a few tens of centimetres of the surface.

DLIS spectra at all altitudes clearly showed an additional signal when the lamp was on. However, at the highest altitudes, the lamp reflection from the surface was negligible, so the additional signal was solely due to scattered light from the lamp into the instrument. This scattered light was estimated from the intensity level in the strong methane bands in the highest-altitude spectrum recorded with the lamp on, and removed from all DLIS spectra. After this correction, only spectra at 36 m and especially 21 m showed significant signal from the lamp. This signal dominated the upward intensity due to solar illumination in all regions of strong and moderate methane absorption, while the latter dominated in the methane windows.

This spectrum, which represents the product of the ground reflectivity and the two-way methane transmission, is shown in Fig. 14 and compared to synthetic spectra with methane mole fractions of 3%, 5% and 7%. The ground reflectivity assumed in

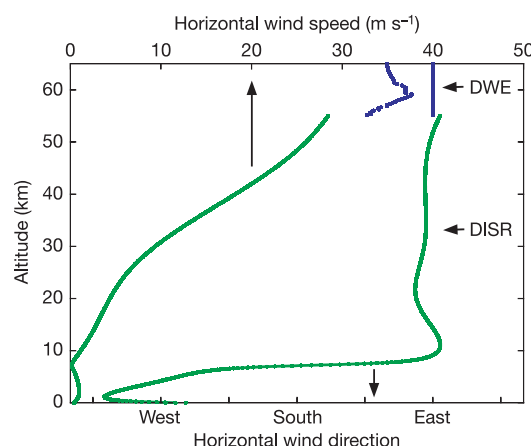


Figure 10 | Observed winds. Horizontal wind speed and direction (counter-clockwise from east) as a function of altitude. The green lines are the DISR data and the blue lines are the high altitude DWE data³³ (showing reasonable consistency between the two). The lines on the left show the wind-speed profile, and on the right is the wind direction. The wind is computed from the combined longitude and latitude reconstructions displayed in Fig. 9. Titan's prograde winds slow from about 28 m s⁻¹ at 50 km to 10 m s⁻¹ near 30 km altitude, then decrease more rapidly from 30 km (10 m s⁻¹) down to 7 km where they drop to zero. Below 7 km (which is near the expected top of the planetary boundary layer) the winds reverse and become retrograde, and the speed increases to about 1 m s⁻¹ around 2–3 km before dropping to almost zero (~ 0.3 m s⁻¹) near the surface. The direction begins as due east, and then turns through south (beginning between 9–7 km) to the west-northwest between 7–5 km. The winds are extrapolated to be retrograde at the surface, but the two-sigma error bars (not shown) of 1 m s⁻¹ at the surface could include surface prograde winds. The error bars at 55 km altitude (4 m s⁻¹) are consistent with continuity from the DWE measurements.

these model calculations is shown in the inset. Four methane bands are seen in the lamp-only spectrum. The good correlation with the models, notably in the weak structures at 1,140, 1,370 and 1,470 nm, indicates a high signal-to-noise ratio of about 50. The best fit is achieved with a methane mole fraction of 5%, which is in firm agreement with the 4.9% *in situ* measurements made by the Gas Chromatograph Mass Spectrometer¹¹. Most structures are well reproduced; notable exceptions are the detailed shape of the 1,000 nm band and the absorption shoulder near 1,320 nm.

We conclude that the methane abundance is $5 \pm 1\%$ in the atmosphere near the surface. The corresponding (two-way, and including the 20° lamp inclination) methane column abundance in the spectrum is 9.6 m-amagat (or a column 9.6 m high at the standard temperature and pressure of 273 K and 1 atmosphere). With a temperature of 93.8 K and pressure of 1467.6 mbar (ref. 12), the relative humidity of methane is about 50% using Brown and Ziegler's saturation law¹³. Therefore, methane near the surface is not near saturation, and ground fogs caused by methane in the neighbourhood of the landing site are unlikely.

The ratio of the observed spectrum to the methane transmission, restricted to spectral regions where the latter is higher than 90%, is shown in Fig. 15a ('plus' symbols). It is compared to one of the DLIS spectra after landing, divided by the lamp spectral response and rescaled by a constant reflectivity factor. Note that although this spectrum shows signs of methane absorption at 1,140–1,160 nm and 1,300–1,400 nm, no attempt of correction was made, in the absence so far of accurate information on the absorption path lengths. The agreement between the shapes of the two independent determinations of the ground reflectivity adds confidence to the result.

The four major characteristics of the surface spectrum are: (1) a relatively low albedo, peaking around 0.18 at 830 nm; (2) a red slope

in the visible range; (3) a quasilinear decrease of the reflectivity by a factor of about two between 830 and 1,420 nm; and (4) a broad absorption, by $\sim 30\%$ of the local continuum, apparently centred near 1,540 nm (although its behaviour beyond 1,600 nm is poorly constrained) as seen in Fig. 15b. This spectrum is very unusual and has no known equivalent on any other object in the Solar System.

Ground-based spectroscopic observations have provided strong evidence, although spectrally restricted to the methane windows, for the presence of water ice on Titan's surface¹⁴, coexisting in variable proportions with a dark component, presumably of organic nature^{15,16}. Water ice may explain the 1,540 nm band, as illustrated in Fig. 15b by a simulation of the reflectance spectrum of a mixture of low-temperature water ice¹⁷, yellow tholins¹⁸ and a spectrally neutral dark component. This identification is reasonable in the context of the light-coloured rocks present at the landing site (Fig. 4), but not conclusive, because some organics do show absorption at a similar wavelength. This is the case, notably, for bright yellow-orange tholins produced in laboratory experiments^{19,20} (shown in Fig. 15b), which partly contribute to this band in the simulation and which may account for the red slope in the visible range of the surface spectrum. It is probably this material, existing as aerosol particles, that absorbs the blue wavelengths, which would explain the yellow-orange colour of Titan's atmosphere as seen from space or from the surface.

We note the remarkable absence of other absorption features in the surface spectrum along with the 1,540 nm band. This is at odds with predictions that some specific chemical bonds, in particular C–H or C $\equiv$ N, and possibly the individual bands of atmospherically abundant species, such as ethane (C₂H₆), acetylene (C₂H₂), propane (C₃H₈), ethylene (C₂H₄), hydrogen cyanide (HCN) and their polymers, would show up as signatures in the surface spectrum.

The most intriguing feature in the surface spectrum is its quasilinear featureless 'blue slope' between 830 and 1,420 nm. As briefly illustrated in Fig. 15b, a featureless blue slope is not matched by any combination of laboratory spectra of ices and complex organics, including various types of tholins. Depending on their composition and structural state (for example, abundance, extension and/or clustering of *sp*² carbon bonds), organic materials in the near-infrared exhibit either distinct absorption bands (for example, bright yellow-orange tholins,

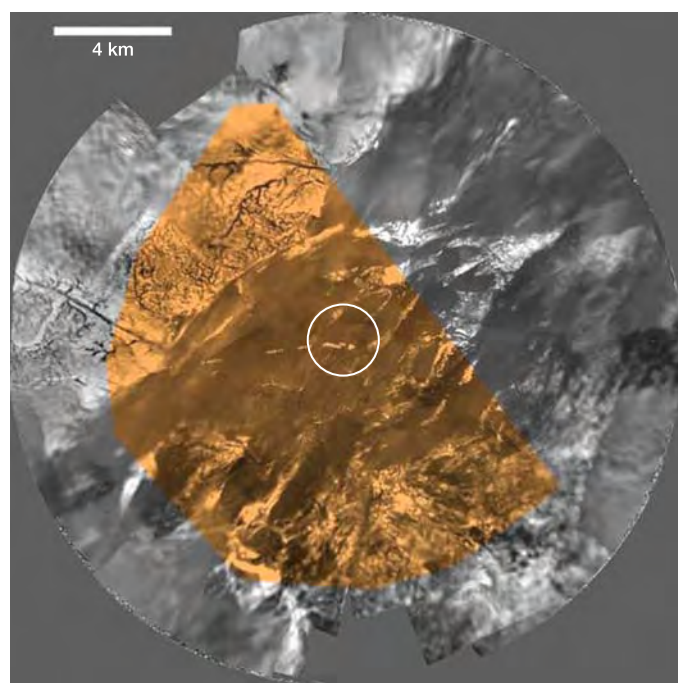


Figure 11 | The surface of Titan displayed in true colour. As seen from an altitude of 8 km. See Fig. 12 for the location of the spectrometer's footprints. Some bright features appear to be overexposed because they are too bright for their colour (the brightness in this image mainly derives from the near-infrared spectrum). True colour is expressed in red-green-blue (RGB) values that are derived by multiplying the spectra with the Commission Internationale de l'Eclairage colour-matching functions (with the 6,500-K correlated colour temperature, the D65 white point). The circle shown is the extent of the lowest panorama (Fig. 3).

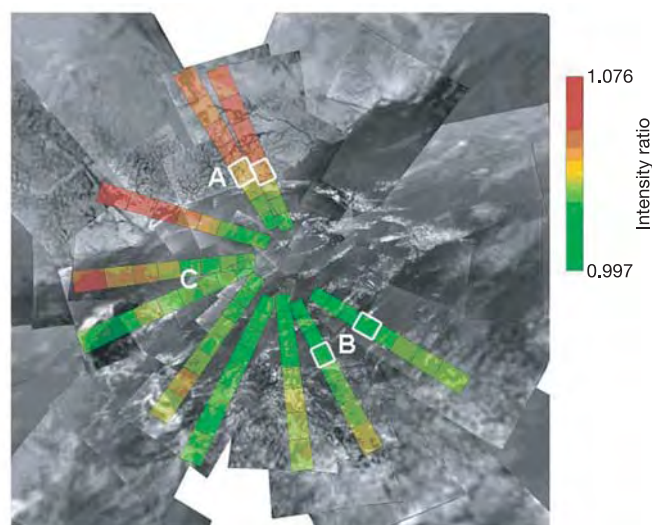


Figure 12 | Reflectivity samples of Titan's surface. A panorama of Titan's surface overlaid with DLVS footprints coloured according to the 827 nm/751 nm intensity ratio, coded from red (high) to green (low). Spectral footprints (the small rectangles) selected for analysis in Fig. 13 are outlined in white. The panorama shows an area of 23 by 23 km. Areas A, B and C are referred to in Fig. 13.

Fig. 15b), or a feature-poor red slope (for medium to low-albedo organics), or a very dark and flat spectrum^{18,21}.

Assessing the material responsible for the blue slope is a major challenge and also a prerequisite for a secure identification of the 1,540 nm band. If this band is indeed mostly due to water ice, an intimate mixing of this ice with a material displaying a strong 'infrared-blue' absorption would explain the absence of the weaker H₂O bands at 1.04 and 1.25 μm in the surface spectrum, as demonstrated for several dark icy satellites, where these bands are hidden by the presence of an organic component (but neutral or reddish). Decreasing the water-ice grain size alone cannot suppress the 1.04- and 1.25- μm bands and at the same time maintain the apparent blue slope that is produced by large-grained water ice (considering only the continuum absorption between the infrared bands). To hide these weak water bands efficiently, the mixture would need to be ice and a material having a stronger and decreasing-with-wavelength infrared absorption.

Haze particle size

The haze particles in Titan's atmosphere have long been known to produce both high linear polarization and strong forward scattering. This has been taken to imply that the particles are aggregates of small 'monomers' in open structures. The amount of linear polarization constrains the size of the small dimension (monomer radius) while the forward scattering or wavelength dependence of extinction optical depth determines the overall size of the particle or the number of monomers constituting the aggregate.

The DISR instrument measured the degree of linear polarization of scattered sunlight by measuring a vertical strip of sky in two bands centred at 492 and 934 nm. Some 50 measurements of this type were collected during the Titan descent. For the small monomer sizes expected, the direction of polarization would be perpendicular to the scattering plane and reach a maximum near 90° scattering angle at an azimuth opposite to the Sun, and would have a maximum electric field vector in the horizontal direction. We eliminated any polarization measurements made by the DISR that did not have this character, assuming that they were not made at the desired azimuth.

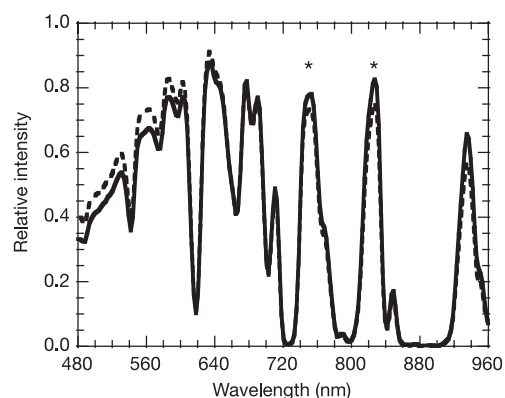


Figure 13 | Spectral comparison of bright highlands and dark lowlands. Spectra of the dendritic river highlands and lakebed lowlands areas are compared. To restrict the influence of the atmosphere, only spectra recorded at the same nadir angle are selected. The solid and dashed curves are the average spectra associated with the two spectral pixels outlined in white near the areas marked 'A' (dendritic highlands, solid) and 'B' (dark lakebed, dashed), respectively in Fig. 12. The two pixels bordering 'C' in Fig. 12 yield a spectrum intermediate to the 'A' and 'B' spectra. The spectra have been corrected for albedo by dividing by the total intensity to emphasize the difference in slope. Not shown here is that the reflectivity of the dendritic area ('A') is higher than that of the lakebed area ('B') at all wavelengths by roughly a factor of two. The asterisks denote the methane windows taken for the reddening ratio in Fig. 12.

Several polarization measurements showing the expected behaviour in Titan's atmosphere were obtained. A gradual rise to a maximum near a scattering angle of 90° was observed, followed by a decrease on the other side of this peak. The solar aureole camera made several of these measurements at different times through the descent that show a smooth decrease in polarization with increasing optical depth into the atmosphere (Fig. 16). Figure 16a shows a maximum degree of linear polarization of about 60% at altitudes above 120 km in the 934-nm channel. Below, we show that the optical depth at 934 nm is a few tenths at this location in the descent. Comparisons of this degree of polarization with model computations for different-sized fractal aggregate particles produced by binary cluster collision aggregation indicate that the radii of the monomers comprising the aggregate particles is near 0.05 μm , almost independent of the number of monomers in the particle.

Haze optical depth and vertical distribution

Before the Huygens probe descent, several workers considered the possibility that the haze in Titan's atmosphere clears below an

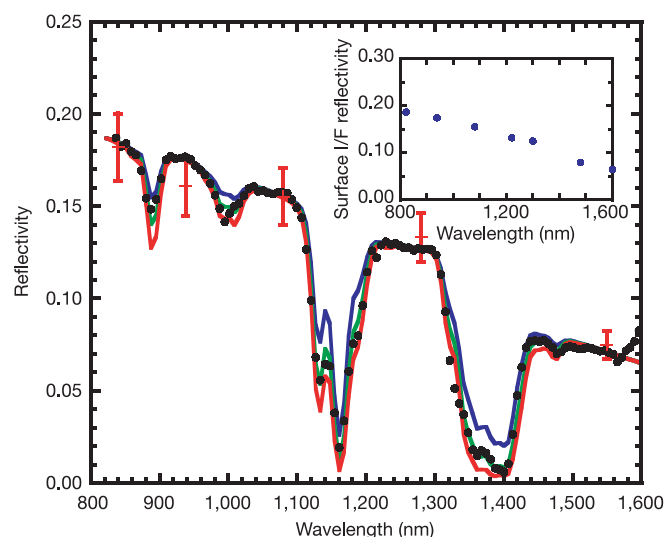


Figure 14 | Derivation of methane mole fraction. Lamp-only downward-looking spectrum from altitude of 21 m (black data points). The red line with three-sigma error bars indicate absolute reflectivity in methane windows estimated from infrared measurements. This spectrum is compared to three models: 3% (blue), 5% (green), and 7% (red) methane mole fractions. These models make use of surface reflectivity at seven wavelengths (shown in inset by the blue points; I/F is the ratio of the intensity to the solar flux divided by π) and linearly interpolated between. From the lamp-on infrared spectra, a lamp-only spectrum at 21 m (representing the spectrum observed by DLIS in the absence of solar illumination) was obtained as follows. First, the reflectivity in regions of negligible methane absorption (at 840, 940, 1070, 1280 and 1,500–1,600 nm) was estimated by the ratio of mean upward intensity (measured by DLIS) to mean downward intensity (measured by ULIS). The mean upward intensity is the average measured over the seven low-altitude DLIS spectra showing no contribution from the lamp (734 to 53 m). The mean downward intensity was obtained by averaging the strongest intensity with the weakest intensity. This average gives reasonable approximation of the downward flux divided by π . The ratio of the mean upward intensity to the mean downward intensity gives reflectivity. Two corrections were required in this analysis: correction for the spatial response of the ULIS diffuser and correction for the solar illumination in the DLIS 21-m spectrum. The correction for diffuser response ranged from 15% (840 nm) to 25% (1,550 nm), assuming a haze optical depth of ~ 2 at 938 nm. The contribution of solar illumination in the DLIS spectrum at 21 m was eliminated by subtracting the average of the DLIS spectra at 85 and 109 m where the lamp contribution was negligible. The difference spectrum was then divided by the spectral response of the lamp and scaled by a constant to match the continuum reflectivities inferred previously, producing the lamp-only spectrum at 21 m.

altitude of some 50 or 70 km (ref. 22) owing to condensation of hydrocarbon gases produced at high altitudes that diffuse to lower, colder, levels of the stratosphere. If such a clearing were to occur in Titan's atmosphere, the intensity seen by the downward-looking DISR spectrometer would be relatively constant below the altitude at which the clearing began.

The brightness looking downward averaged along the DLVS slit and averaged over azimuth increases by a factor of two from the surface to 30 km altitude at a wavelength of 830 nm as shown in Fig. 17a. The increase at 830 nm is due almost solely to scattering by haze between 30 km and the surface. These observations demonstrate that there is significant haze opacity at all altitudes throughout the descent, extending all the way down to the surface.

The brightness of the visible spectra looking upward depends on the azimuth relative to the Sun. Although the probe attitude is not yet well known, it is clear that the minimum intensities are found looking away from the Sun. The upward-looking spectra looking away from the Sun start with low intensities at the highest altitudes and increase in intensity as the altitude decreases from 140 to about 50 km (see Fig. 17b). Below 50 km the intensity decreases at short wavelengths as altitude decreases, while the intensity in continuum regions longer than 700 nm continues to increase, as shown in Fig. 17c.

The intensity looking upward away from the Sun, and the azimuthally averaged intensity looking downward at each continuum wavelength as functions of altitude, constrain the vertical distribution of aerosol opacity in the atmosphere as well as the aerosol single-scattering albedo. With the monomer radius fixed at

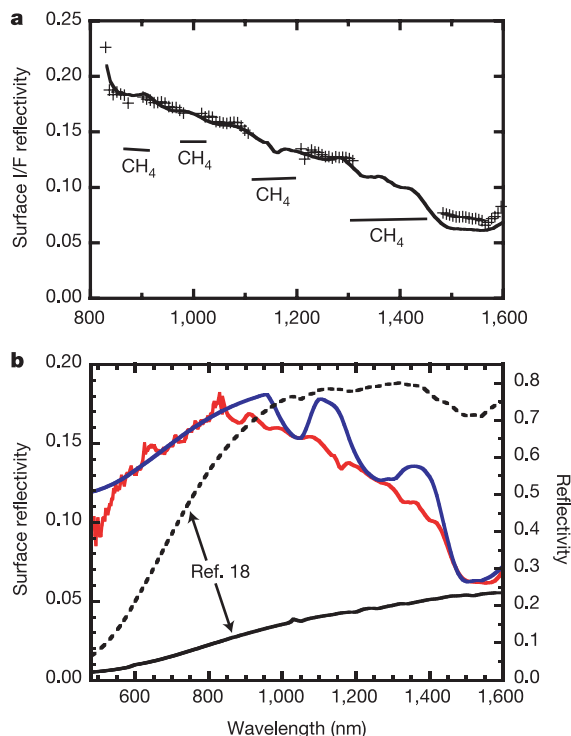


Figure 15 | Reflectance of Titan's surface. **a**, Reflectivity measured from 21 m altitude ('plus' symbols) compared to the reflectivity after landing (solid curve). The methane absorption bands are indicated by the CH₄ symbol. **b**, Surface reflectivity as measured after landing (red line). It is compared with a simulation (blue line) of a mixture of large-grained (750 μ m) low-temperature water ice, yellow tholins, and an unknown component with featureless blue slope between 850 and 1,500 nm. Spectra of two different organic tholins: a yellow tholin (dashed line) and a dark tholin (solid black line) from ref. 18 are also shown for comparison (reflectance scale reduced by a factor of 4). We are attempting to identify or synthesize the missing blue material in our laboratory.

0.05 μ m from the polarization measurements, the adjustable parameters include the number of monomers in each aggregate particle, N , as well as the local particle number density, n , in cm^{-3} as a function of altitude. An algorithm developed by (and available from) M.L. was used to determine the single-scattering phase function, the single-scattering albedo, and the extinction cross-section for each aggregate particle as functions of the wavelength, the real and imaginary refractive indices, the monomer radius, and the number of monomers per aggregate particle. This algorithm is based on the discrete dipole approximation and the T-matrix method (M. Lemmon, personal communication) to evaluate the single-scattering properties of the aggregate particles. These computations are most accurate at relatively small particle sizes and depend on extrapolation for N of 256 or larger.

For large particles the wavelength dependence of the extinction optical depth is smaller than for small particles. An N larger than about 100 is required to fit the observations. Hence, models with $N = 256$ or 512 monomers per particle are shown, even though for these values of N the single-scattering algorithm is not as accurate as desired. For these initial models, the real and imaginary refractive indices for the aerosols are taken from the measurements of laboratory tholins in ref. 23.

The radius, R_p , of the circle having the same projected area as an aggregate particle is given by $R_p = r\sqrt{N^{0.925}}$, where r is the monomer radius and N is the number of monomers. Particles with 256 or 512 monomers have the same projected areas (which control their forward scattering properties) as circles with radii 0.65 and 0.9 μ m, respectively.

Comparison of the observed downward-streaming intensity looking away from the Sun at wavelengths of 531 and 829 nm with

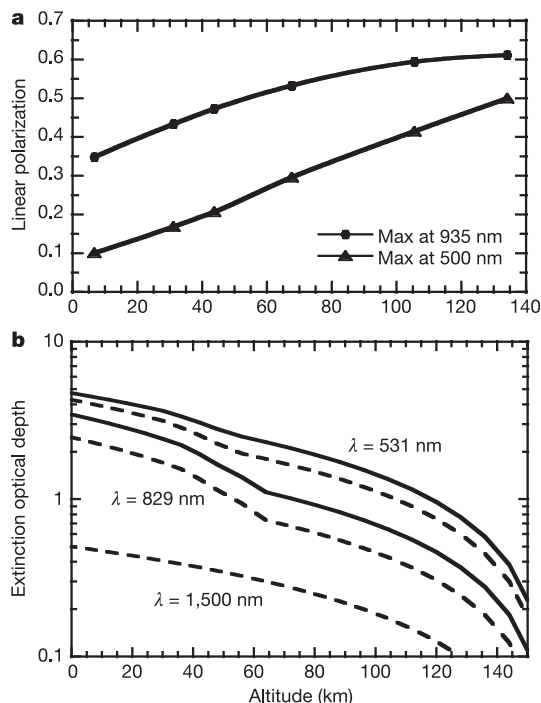


Figure 16 | Haze properties. **a**, The maximum degree of linear polarization measured opposite to the Sun as a function of altitude in our 500-nm channel (triangles) and in the 935-nm channel (dots). **b**, Extinction optical depth versus altitude for three wavelengths, 531 (top), 829 (middle) and 1,500 nm (bottom). The dashed curves correspond to $N = 256$ monomers, and the solid curves correspond to $N = 512$ monomers of the aggregate particles that make up Titan's haze. Note that the 531 (top) curve was constrained above 40 km and extrapolated to the ground. More explicit constraints from the infrared spectrometer will be available after the probe azimuth with time is determined.

plane-parallel radiative transfer models constrains the vertical distribution of optical depth in Titan's atmosphere. The vertical distribution of particles can be adjusted to fit these curves arbitrarily well. In this preliminary work, only one constant number density

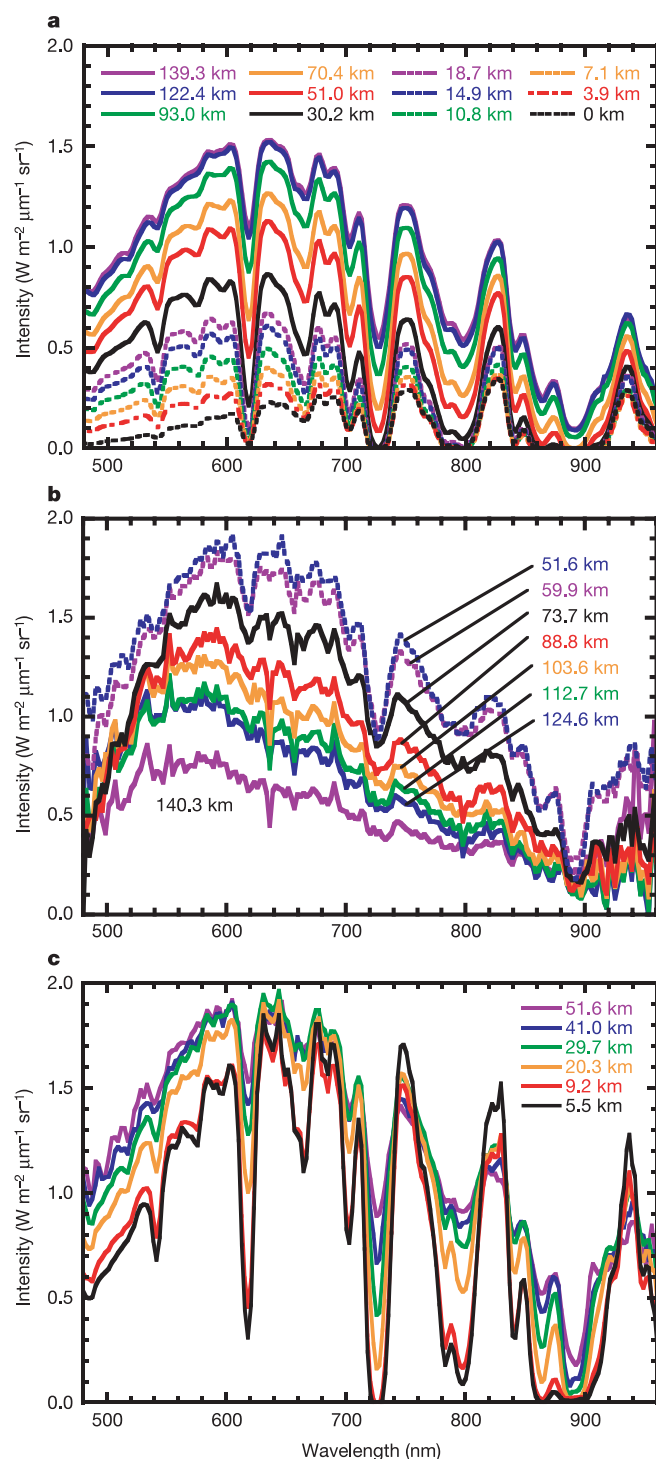


Figure 17 | Atmospheric spectra. **a**, The average intensity looking downward averaged over azimuth and over the length of the slit (10° to 50° nadir angle) as a function of wavelength for several altitudes as labelled. **b**, The intensity measured by the Upward-Looking Visible Spectrometer in the direction opposite the Sun as a function of wavelength for several altitudes as labelled. Note that the brightness begins at a low level at 140 km, and increases as altitude decreases. **c**, Same as **b** but for altitudes below 50 km. Note that the brightness away from the Sun decreases with decreasing altitude at short wavelengths, but increases in continuum regions longward of 700 nm.

above an altitude of 80 km and a second constant number density below 80 km were considered for models with $N = 256$ and $N = 512$, as shown in Fig. 18a. The number densities are larger in the lower half of the atmosphere than the upper half, but only by modest factors of two to three. The number densities are not exactly equal in the models at different wavelengths, but this is probably due to the extrapolation in the wavelength dependence of the cross-sections in the models for these relatively large N values at the shortest wavelengths (and largest size parameters). Average number densities in the entire atmosphere between 30 and 65 cm^{-3} are required if the number of monomers per particle is 256. Average number densities between 15 and 40 cm^{-3} are required if N is 512.

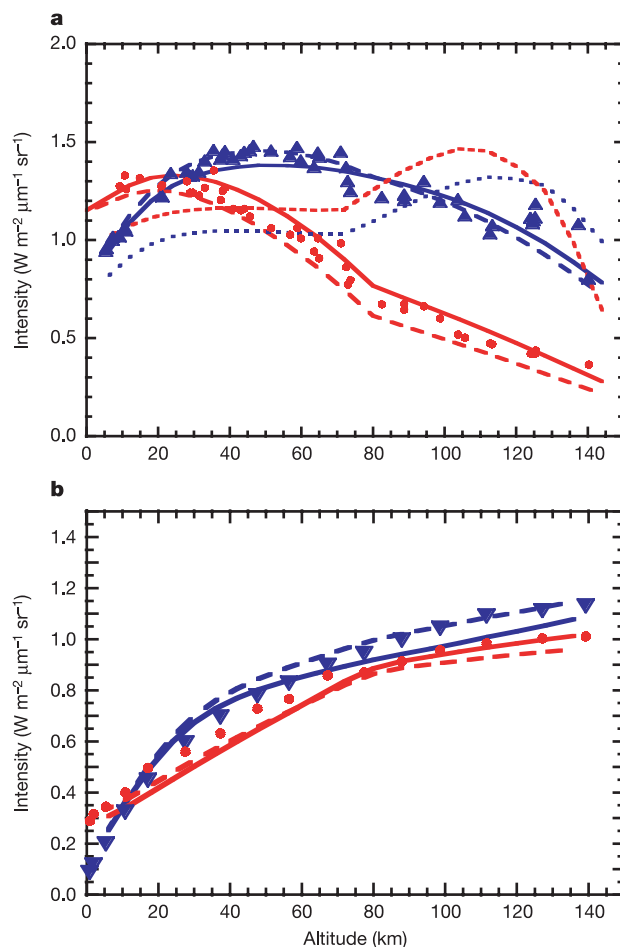


Figure 18 | Haze models versus observations. **a**, Measured upward-looking intensity (points) away from the Sun versus altitude for 531 (blue) and 829 nm (red). Three models are shown compared to the observations at each wavelength. The solid curves are for 512 monomers in each aggregate particle. The model at 531 nm has 12 particles cm^{-3} above 80 km and 18 particles cm^{-3} below that altitude. The corresponding model at 829 nm has 20 particles cm^{-3} above 80 km and 60 cm^{-3} below. The models indicated by long-dashed lines have 256 monomers per particle and at 531 nm the number density is 20 particles cm^{-3} above 80 km and 40 cm^{-3} below. At 829 nm the number density is 30 cm^{-3} above 80 km and 100 cm^{-3} below. The number density of particles differs slightly with wavelength because the model of fractal aggregate particles does not yet reproduce the wavelength dependence of the cross-section to high accuracy. The models indicated by short-dashed lines have 256 monomers per particle, and have the same number of total particles as the models indicated by long-dashed lines, but all the particles are concentrated above 72 km with a clear space below. Such models with clear spaces are clearly not in agreement with the observations. **b**, Downward-looking measured intensities versus altitude (plotted as points) for 531 (blue points) and 829 nm (red points). The two models (plotted as curves) are the same models as those shown by long-dashed lines and solid curves in **a**.

Models with a clear space below an altitude of 72 km are also shown in Fig. 18a. It is apparent that no such clear space exists in the region of the probe's entry. It will be interesting to examine the range of parameters in cloud physics models needed to reproduce the continuous variation of haze opacity throughout Titan's atmosphere.

Haze models must reproduce the upward-streaming intensity observed in the atmosphere as well as the downward-streaming intensity. Models with $N = 256$ and 512 at 531 and 829 nm are compared to the upward intensity averaged in azimuth and along the slit in Fig. 18b. While the fit is not exact, it is clear that the models that fit the downward intensity away from the Sun are also in reasonable agreement with the measured upward intensities. It is interesting to note the ground reflectivity implied by the measurements at 531 and 829 nm. These values include the true shape of the diffuser in the Upward-Looking Visible Spectrometer and produce ground reflectivities of 0.13 at 531 nm and 0.19 at 829 nm. The corrected value at 829 nm is in good agreement with the value measured by the Infrared Spectrometer (0.18) when a correction for the diffuser's non-ideal shape is made to the infrared spectrometer measurements.

Haze structure and methane absorptions

Haze models must fit the observations at all wavelengths. How well do the models derived from the visible spectrometer fit the DISR observations in the infrared? The ULIS spectra in Fig. 19 clearly show absorption by the methane bands around 890, 1,000, 1,160 and

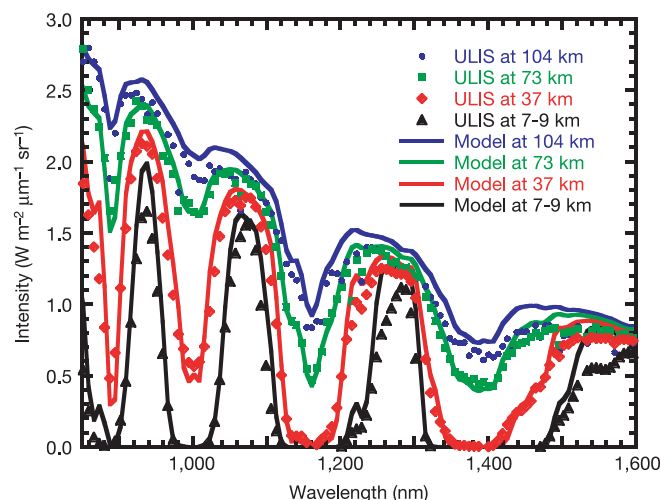


Figure 19 | Vanishing sunlight. ULIS spectra (points) recorded at various altitudes and showing the growth of the methane-band absorption with depth in the atmosphere. The spectra at altitudes greater than 3.5 km have been integrated over several probe rotations and correspond approximately (but not exactly) to azimuth-averaged intensities. The full analysis of these observations must await refinement of the attitude of the probe as a function of time, a task still in progress. Models, with a methane mole fraction of 1.6% in the stratosphere increasing to 5% at the surface, are shown for comparison (lines). At low altitudes (<20 km), the mismatch between model and observations in the methane windows, specifically around 1,280 and 1,520–1,600 nm, is probably due to errors in the methane absorption coefficients at long path-lengths. The model slightly overestimates the intensity in the 104-km spectrum, because the latter does not correspond to an exact azimuth average. Radiative transfer calculations were based on a 16-term exponential-sum formulation of the methane absorption properties. In the near-infrared ($\lambda > 1,050$ nm), these absorption coefficients were calculated from a band model with a modified temperature dependence designed to better match the low-temperature observations. In the visible ($\lambda < 1,050$ nm), the absorption coefficients of Karkoschka³⁴ were used. In practice, for 30 pressure–temperature conditions representative of 30 levels in Titan's atmosphere¹², methane transmissions were calculated for 60 different paths, convolved to the resolution of the DISR spectrometers, and this ensemble of convolved transmissions was fitted each time with an exponential-sum model.

1,380 nm. The depths of these bands increase with decreasing altitude as a result of increasing methane column density. They are correctly reproduced by radiative transfer calculations based on an exponential sum formulation for the methane absorption and a stratospheric methane mole fraction of 1.6% (ref. 24). The agreement is worse at low altitudes in the troposphere, probably owing to inaccuracies in the methane absorption coefficients. In the methane windows the downward average intensity varies by 25% or less between 104 and 10 km, indicating relatively low aerosol absorption in the infrared range.

At altitudes less than 3.5 km, we used single short exposures for the infrared spectra rather than long time averages. The 940 nm intensity in the last three ULIS spectra is about four times larger than in the first three spectra, indicating that the Sun is located in their field of view (see Fig. 20). The contrast between the most intense spectrum (with the Sun in the field of view) and the weakest one (with the Sun out of the field of view) increases with wavelength, reaching 17 at 1,550 nm, a consequence of the decreasing haze optical depth. This contrast can be used to constrain the haze optical depth, assuming that the spectra correspond approximately to solar azimuths of 0 and 180°. A satisfactory model, using aggregate particles of 256 monomers, a 0.05- μm monomer radius, and a uniform concentration of 52 particles cm^{-3} , indicates an optical depth of about 2 at 940 nm, decreasing to 0.5 at 1,550 nm. Models with one-half and twice the particle density (and hence optical depth) yield a contrast between spectra with the Sun in and out of the field about twice as large and half as large (respectively) as observed.

The methane bands are prominent in the DLIS spectra at all altitudes (see Fig. 21a). The residual intensity in the cores of these bands is due to scattering by aerosol particles between the probe and the surface. Its variation provides a constraint of the vertical profile of the haze particles between approximately 150 and 40 km, as illustrated in Fig. 21a. The method is not sensitive at low altitudes because of absorption of the downward solar flux in the methane bands. A model with a constant particle concentration with altitude provides a

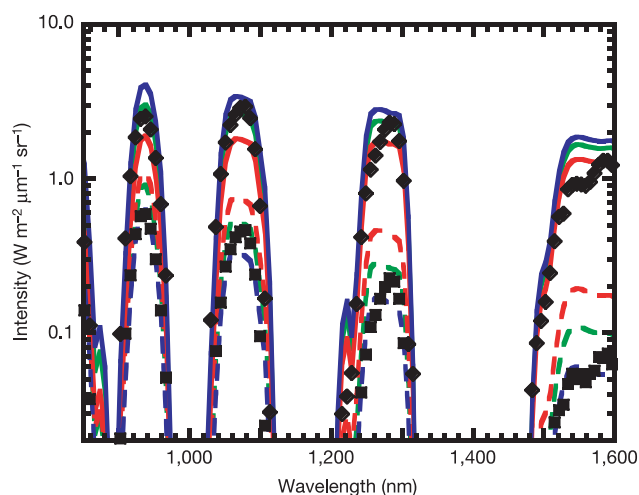


Figure 20 | Determination of total haze optical depth. ULIS spectra (black points) recorded at 734 m (diamonds) and 53 m (squares) above the surface with integration times of 1 s. The one with the highest intensity (734 m) has the Sun in its field of view; the lower one (53 m) does not. The contrast between the two in the methane windows increases with wavelength and is a sensitive function of the haze optical depth. The nominal model, shown for comparison (green line), has an optical depth of 2 at 940 nm decreasing to 0.5 at 1,550 nm. Calculations correspond to intensities averaged over the field of view and azimuths of 0 and 180 degrees with respect to the Sun. Other models show the effect of doubling (red) and halving (blue) the particle concentration. Solid lines show model intensity towards the Sun while the dashed lines show the intensity with the instrument facing away from the Sun.

good fit of the methane bands. Moderate variations of the particle concentration with height are also acceptable, but a model with a clear space in the lower stratosphere is inconsistent with the data.

The optical depths in the models that fit the visible and infrared spectral observations with $N = 256$ and 512 are shown as functions of altitude in Fig. 16b. The number density was assumed to be 52 cm^{-3} , independent of altitude for the infrared models with $N = 256$. The models computed for comparison with the visible spectrometer contained constant and different number densities above and below 80 km. The average number density above and below 80 km for the models with $N = 256$ (30 and 65) are in reasonable agreement with the single value used in the models derived from the infrared spectrometer. The same particle number densities give the required optical depths from 900 to 1,550 nm, indicating that the algorithm for generating cross-sections from particle sizes is working in a consistent manner. At shorter wavelengths (531 nm), the size parameter is sufficiently large that the

cross-section algorithm is not as accurate, and the number density is decreased slightly to give models that fit the observations. The variation of optical depth with wavelength is modest, decreasing by only about a factor of 2.8 from 500 to 1,000 nm. If 512 monomers are used for the particles, the wavelength dependence is even less steep. The haze optical depth as a function of wavelength is presented in Fig. 21b.

A thin layer of haze near 21 km altitude

Many workers have suggested that hydrocarbons produced at very high altitudes could diffuse downward to cooler levels where they could condense on haze particles. Do our intensity profiles looking towards the horizon detect any thin haze layers at specific altitudes that might be due to this mechanism?

Figure 22 shows the normalized profile of intensity measured by the Side-Looking Imager (SLI) compared to a model. The left-hand side of the plot shows normalized intensity as a function of nadir angle for the observations at altitudes ranging from 20.4 to 22.3 km. The observations at 22.1 km and above and at 20.4 km and below show smooth functions of nadir angle. However, for the measurement at 20.8, and the two measurements at 20.9 km altitude, a dip of about 2% is seen near a nadir angle of 90° .

The curves in the right-hand side of Fig. 22 show the intensities of a model having a thin additional layer of haze at an altitude of 20.9 km. The haze layer of vertical absorption optical depth within a factor of two of 0.001 with a gaussian profile between 1 and 2 km thick can reproduce the depth of the feature. The location of the layer is at $21.0 \pm 0.5 \text{ km}$, where the local temperature is 76 K and the pressure is 450 mbar (ref. 12).

This feature at 21 km occurs in the troposphere and may be an indication of methane condensation. It is the only indication of a thin layer seen in the set of SLI images taken from 150 km to the surface. Evidence of condensation of hydrocarbons in the lower stratosphere, where several hydrocarbons might be expected to condense²⁵, has not yet been found, but the search is continuing.

Unravelling Titan's mysteries

Some of the major questions about Titan concern the nature of the source of methane that replaces the irreversible loss at high

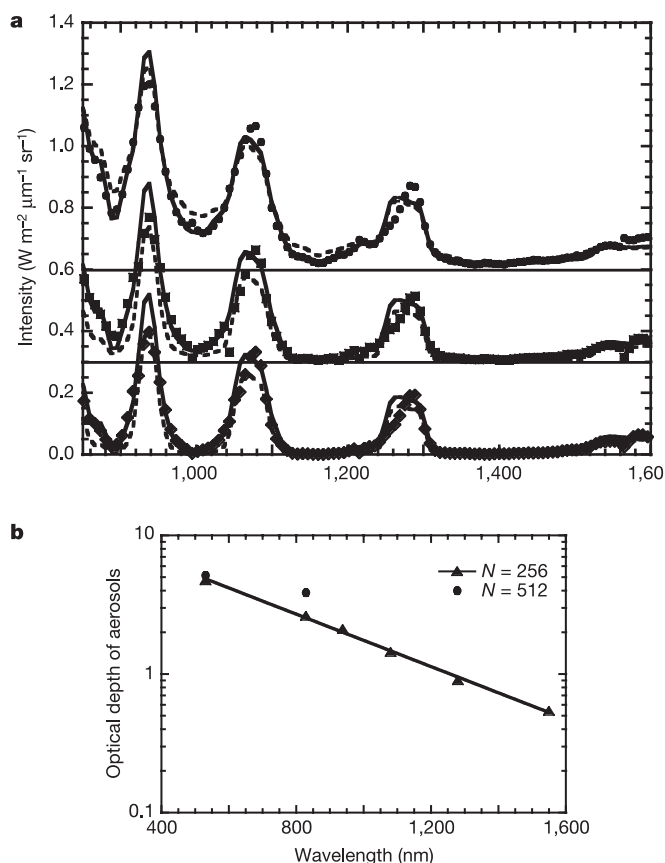


Figure 21 | Haze vertical structure and total optical depth. **a**, DLIS spectra recorded at three altitudes: 104 km (top, dots), 82 km (middle, squares), and 57 km (bottom, diamonds). The intensities for the higher altitudes have been displaced in 0.3 increments for clarity. The points are measured data; the lines are the nominal model; the dashed lines are a modified model with the same optical depth as the nominal model, but with all the haze particles concentrated above 72 km (clearing below). The residual intensity in the core of the methane bands is a sensitive indicator of the presence of scattering particles beneath the probe. The model with the clearing produces too much emission in the core of the CH_4 bands at high altitude and not enough at low altitude. **b**, Total extinction optical depth of the haze alone versus wavelength. The triangles are for models with 256 monomers per particle. The points at the two shortest wavelengths are from models that fit the visible spectrometer measurements. The other four points are from models that fit the infrared spectrometer measurements. The dots are for models with 512 monomers per particle that fit the visible spectrometer measurements.

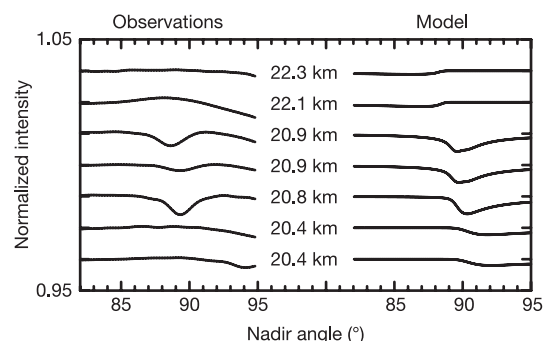


Figure 22 | Thin cloud layer observation at 21 km. In the left-hand side are the intensity profiles as a function of nadir angle divided by the average intensity profile measured by the SLI at the altitudes indicated. The right-hand side of the figure shows the model intensity profiles computed for a cloud layer of absorption optical depth 0.001 which is 1 km thick at an altitude of 21 km. The model is able to reproduce the 2% contrast feature seen in the observations at altitudes of 20.8 and 20.9 km. If the layer is mostly illuminated by diffuse light, the absorption optical depth is equal to total optical depth times the difference between the single-scattering albedos of the material in the layer and the albedo of the background haze. If the layer is primarily illuminated by direct sunlight, the absorption optical depth of the haze is proportional to the total optical depth of the haze times the difference between the phase functions of the material in the layer and the background haze at the scattering angles for any observation.

altitudes by photochemistry that produces a host of complex organic compounds. Open pools of liquid hydrocarbons on the surface have been suggested, as well as cryovolcanism. Also, if methane photochemistry has been occurring over the lifetime of the Solar System the organic products of these processes should have accumulated to significant depths on the surface and should be seen in images and spectra of the surface.

Although no such liquid bodies were directly imaged by DISR, there is compelling evidence for fluid flow on the surface of Titan, including the dendritic and stubby drainage channel networks, the rounded and size-graded 'rocks' at the surface landing site, and the morphology of the shoreline, offshore structures, and the appearance of the darker lakebed region. The stubby networks may imply sapping or spring-fed flows, as the existence of liquid pools on the surface and the frequency of precipitation that could cause the deep dendritic drainage channels are both still unconfirmed. In addition, there are at least a few structures that suggest cryovolcanic flows on the surface.

The ground track derived by the image correlations demonstrates a zonal wind field that is mostly prograde. The general altitude profile and shape agree with predicted average models of the zonal wind flow between 50 and 10 km altitude⁷, although with a reduced intensity. Below 10 km, falling wind speeds and an abrupt change of wind direction indicate a planetary boundary layer some 7–8 km thick, scaling nicely from near-equatorial terrestrial boundary layers.

Spatially resolved spectral reflectance measurements of different regions on the surface suggest that the uplands are redder than the lowland lakebed regions. The regions near the mouths of the rivers are also redder than the lake regions. A host of questions about the sequence of flooding and formation of these structures is suggested by these observations.

The reflectivity of the surface at the landing site was measured from 480 nm to 1,600 nm without the interference of methane absorption bands or haze opacity. The peak reflectivity in the dark regions is about 0.18 at 830 nm and decreases towards longer and shorter wavelengths. The red slope in the visible is consistent with organic material, such as tholins, but the blue infrared slope is still unexplained. Between 1,500 and 1,600 nm the reflectivity is low (0.06) and flat, consistent with water ice. Nevertheless, the decrease in reflectivity from 900 to 1,500 nm does not show the expected weak absorption bands of water ice near 1,000 and 1,200 nm, and the identity of the surface component responsible for this blue slope remains unknown.

The nature of the haze aerosols measured by DISR is different in significant ways from the view before the Huygens mission. Before the Huygens probe, cloud physics models with sedimentation and coagulation predicted a strong increase in haze density with decreasing altitude⁹. In addition, measurements of the high degree of linear polarization in light scattered from Titan near a phase angle of 90° by the Pioneer and Voyager spacecraft could only be matched by spherical particles having radii less than or equal to 0.1 µm. Such small particles produced a strong increase in optical depth with decreasing wavelength shortward of 1,000 nm. Fitting the strong methane band at 890 nm constrained the amount of haze at high altitudes. This haze became optically much thicker at the wavelength of the weaker methane band at 619 nm. To fit the observed strength of this band it was necessary to remove the haze permitted by the cloud physics calculations at altitudes below about 70 km by invoking condensation of organic gases produced at very high altitudes as they diffused down to colder levels²⁶. The condensation of many organic gases produced by photochemistry at high altitudes on Titan was suggested by Sagan and Thompson in ref. 25, and seemed consistent with this view.

The next development in Titan haze models (pioneered by R.W., P.S., Cabane, M. and M.L.) included the use of fractal aggregate haze particles that had a small component (monomer) with a radius of about 0.06 µm to produce strong linear polarization^{26–30}. These

monomers stuck together in an aggregation of many tens (or more) monomers. The large size of the aggregation could produce the strong forward scattering required from the Titan haze aerosols while preserving the high degree of linear polarization. However, it was quite laborious to compute the single-scattering properties of such aggregate particles for more than about 100 monomers at visible wavelengths. Particles with an effective radius of about 0.35 µm were required to produce the degree of forward scattering observed by Voyager³¹. This required the number of monomers in an aggregate particle to be about 45, and permitted single-scattering computations of the cross-section and phase function of the particles over the visible range. Of course, even larger numbers of monomers per particle would have matched the observations at high phase angles on Voyager, but these were difficult to perform and have largely gone unexplored. If larger particles had been used, however, the optical depth of the aerosols at shorter wavelengths would not have been nearly so large, and the clear space below 70 km may well not have been necessary.

The new DISR observations give a measurement of the monomer radius of 0.05 µm, in good agreement with previous estimates. Significantly, however, they show that the haze optical depth varies from about 2 at 935 nm to only about 4.5 at 531 nm, and the number of monomers in a haze particle is therefore probably several hundred. A value of 256 for N gives a projected area equal to that of a sphere of radius 0.65 µm, about twice as large as previously assumed. With $N = 512$, the equivalent sphere with the same projected area has a radius of 0.9 µm, nearly three times the size previously used. In any case, it seems that the size of the aggregate particles is several times as large as in some of the older models. A better estimate of the particle size will be available after the analysis of the solar aureole measurements of the variation in brightness near the Sun. In addition, measurements by the DISR violet photometer will extend the optical measurements of the haze to wavelengths as short as the band from 350 to 480 nm, also helping to constrain the size of the haze particles.

The number density of the haze particles does not increase with depth nearly as dramatically as predicted by the older cloud physics models. In fact, the number density increases by only a factor of a few over the altitude range from 150 km to the surface. This implies that vertical mixing is much less than had been assumed in the older models where the particles are distributed approximately as the gas is with altitude. In any case, no clear space at low altitudes, which was suggested earlier³², was seen.

The methane mole fraction of 1.6% measured in the stratosphere by the Composite Infrared Spectrometer (CIRS) and the Gas Chromatograph Mass Spectrometer is consistent with the DISR spectral measurements. At very low altitudes (20 m) DISR measured $5 \pm 1\%$ for the methane mole fraction.

Finally, the entire set of DISR observations gives a new view of Titan, and reinforces the view that processes on Titan's surface are more similar to those on the surface of the Earth than anywhere else in the Solar System.

Received 26 May; accepted 8 August 2005.

Published online 30 November 2005.

1. Coustenis, A. *et al.* Maps of Titan's surface from 1 to 2.5 µm. *Icarus* **177**, 89–105 (2005).
2. Porco, C. C. *et al.* Imaging of Titan from the Cassini spacecraft. *Nature* **434**, 159–168 (2005).
3. Sotin, C. *et al.* Infrared images of Titan. *Nature* **435**, 786–789 (2005).
4. Elachi, C. *et al.* Cassini radar views the surface of Titan. *Science* **308**, 970–974 (2005).
5. Tomasko, M. G. *et al.* The Descent Imager/Spectral Radiometer (DISR) experiment on the Huygens entry probe of Titan. *Space Sci. Rev.* **104**, 469–551 (2002).
6. Ivanov, B. A., Basilevski, A. T. & Neukem, G. Atmospheric entry of large meteoroids: implication to Titan. *Planet. Space Sci.* **45**, 993–1007 (1997).
7. Flasar, F. M., Allison, M. D. & Lunine, J. I. Titan zonal wind model. *ESA Publ. SP-1177*, 287–298 (1997).

8. Bond, N. A. Observations of planetary boundary-layer structure in the eastern equatorial Pacific. *J. Atmos. Sci.* **5**, 699–706 (1992).
9. Lindal, G. F., Wood, G. E., Hotz, H. B. & Sweetnam, D. N. The atmosphere of Titan: An analysis of the Voyager 1 radio occultation measurements. *Icarus* **53**, 348–363 (1983).
10. Tokano, T. & Neubauer, F. M. Tidal winds on Titan caused by Saturn. *Icarus* **158**, 499–515 (2002).
11. Niemann, H. B. *et al.* The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe. *Nature* doi:10.1038/nature04122 (this issue).
12. Fulchignoni, M. *et al.* *In situ* measurements of the physical characteristics of Titan's environment. *Nature* doi:10.1038/nature04314 (this issue).
13. Brown, G. N. Jr & Ziegler, W. T. in *Advances in Cryogenic Engineering* (ed. Timmerhaus, K. D.) Vol. 25, 662–670 (Plenum, New York, 1980).
14. Griffith, C. A., Owen, T., Geballe, T. R., Rayner, J. & Rannou, P. Evidence for the exposure of water ice on Titan's surface. *Science* **300**, 628–630 (2003).
15. Coustenis, A., Lellouch, E., Maillard, J.-P. & McKay, C. P. Titan's surface: composition and variability from the near-infrared albedo. *Icarus* **118**, 87–104 (1995).
16. Lellouch, E., Schmitt, B., Coustenis, A. & Cuby, J.-G. Titan's 5- μ m lightcurve. *Icarus* **168**, 204–209 (2004).
17. Grundy, W. & Schmitt, B. The temperature-dependent near-infrared absorption spectrum of hexagonal H₂O ice. *J. Geophys. Res.* **E 103**, 25809–25822 (1998).
18. Bernard, J.-M. *et al.* Evidence for chemical variations at the micrometric scale of Titan's tholins: Implications for analysing Cassini-Huygens data. *Icarus* (submitted).
19. Coll, P. *et al.* Experimental laboratory simulation of Titan's atmosphere: aerosols and gas phase. *Planet. Space Sci.* **47**, 1331–1340 (1999).
20. Bernard, J.-M. *et al.* Experimental simulation of Titan's atmosphere: detection of ammonia and ethylene oxide. *Planet. Space Sci.* **51**, 1003–1011 (2003).
21. Moroz, L. V., Arnold, G., Korochantsev, A. V. & Wäsch, R. Natural solid bitumens as possible analogs for cometary and asteroid organics. 1. Reflectance spectroscopy of pure bitumens. *Icarus* **134**, 253–268 (1998).
22. Toon, O. B., McKay, C. P., Griffith, C. A. & Turco, R. P. A physical model of Titan's aerosols. *Icarus* **95**, 24–53 (1992).
23. Khare, B. N. *et al.* Optical constants of organic tholins produced in a simulated Titanian atmosphere—From soft X-ray to microwave frequencies. *Icarus* **60**, 127–137 (1984).
24. Flasar, F. M. *et al.* Titan's atmospheric temperatures, winds, and composition. *Science* **308**, 975–978 (2005).
25. Sagan, C. & Thompson, W. R. Production and condensation of organic gases in the atmosphere of Titan. *Icarus* **59**, 133–161 (1984).
26. Lemmon, M. T. *Properties of Titan's Haze and Surface*. PhD dissertation, (Univ. Arizona, 1994).
27. West, R. A. & Smith, P. H. Evidence for aggregate particles in the atmospheres of Titan and Jupiter. *Icarus* **90**, 330–333 (1991).
28. Cabane, M., Chassefière, E. & Israel, G. Formation and growth of photochemical aerosols in Titan's atmosphere. *Icarus* **96**, 176–189 (1992).
29. Cabane, M., Rannou, P., Chassefière, E. & Israel, G. Fractal aggregates in Titan's atmosphere. *Planet. Space Sci.* **41**, 257–267 (1993).
30. West, R. A. Optical properties of aggregate particles whose outer diameter is comparable to the wavelength. *Appl. Opt.* **30**, 5316–5324 (1991).
31. Rages, K. B. & Pollack, J. Vertical distribution of scattering hazes in Titan's upper atmosphere. *Icarus* **55**, 50–62 (1983).
32. McKay, C. P. *et al.* Physical properties of the organic aerosols and clouds on Titan. *Planet. Space Sci.* **49**, 79–99 (2001).
33. Bird, M. K. *et al.* The vertical profile of winds on Titan. *Nature* doi:10.1038/nature04060 (this issue).
34. Karkoschka, E. Methane, ammonia, and temperature measurements of the Jovian planets and Titan from CCD-spectrophotometry. *Icarus* **133**, 134–146 (1998).

Acknowledgements We thank the people from the following organizations whose dedication and effort have made this project successful: AETA (Fontenay-aux-Roses, France), Alcatel Space (Cannes, France), Collimated Holes Inc., EADS Deutschland GmbH (formerly Deutsche Aerospace AG, Munich, Germany), ETEL Motion Technology (Mortiers, Switzerland), The European Space Agency's (ESA) European Space and Technology Centre (ESTEC), The European Space Operations Centre (ESOC), The Jet Propulsion Laboratory (JPL), Laboratoire de Planétologie de Grenoble (CNRS-UJF), Loral Fairchild (Tustin, California, USA), Martin Marietta Corporation (Denver, Colorado, USA), Max-Planck-Institut für Sonnensystemforschung (Katlenburg-Lindau, Germany), Observatoire de Paris (Meudon, France), Technische Universität Braunschweig (TUB), Thomson-CSF (Grenoble, France), University of Arizona's Kuiper Lunar and Planetary Laboratory (LPL), and the US Geological Survey (Flagstaff, Arizona, USA).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.S. (csee@lpl.arizona.edu).

The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe

H. B. Niemann¹, S. K. Atreya², S. J. Bauer³, G. R. Carignan², J. E. Demick¹, R. L. Frost⁷, D. Gautier⁴, J. A. Haberman¹, D. N. Harpold¹, D. M. Hunten⁵, G. Israel⁶, J. I. Lunine⁵, W. T. Kasprzak¹, T. C. Owen⁸, M. Paulkovich¹, F. Raulin⁹, E. Raean¹ & S. H. Way¹

Saturn's largest moon, Titan, remains an enigma, explored only by remote sensing from Earth, and by the Voyager and Cassini spacecraft. The most puzzling aspects include the origin of the molecular nitrogen and methane in its atmosphere, and the mechanism(s) by which methane is maintained in the face of rapid destruction by photolysis. The Huygens probe, launched from the Cassini spacecraft, has made the first direct observations of the satellite's surface and lower atmosphere. Here we report direct atmospheric measurements from the Gas Chromatograph Mass Spectrometer (GCMS), including altitude profiles of the constituents, isotopic ratios and trace species (including organic compounds). The primary constituents were confirmed to be nitrogen and methane. Noble gases other than argon were not detected. The argon includes primordial ³⁶Ar, and the radiogenic isotope ⁴⁰Ar, providing an important constraint on the outgassing history of Titan. Trace organic species, including cyanogen and ethane, were found in surface measurements.

Determining the composition of the atmosphere of Titan and the nature of the aerosols making up the surface-hiding haze layers are two of the primary objectives of the Cassini-Huygens mission. Whereas nitrogen (N₂) and methane (CH₄) were well established as the major atmospheric constituents after the Voyager 1 encounter¹, the vertical distribution of methane, the isotopic ratio of N in N₂ and the abundances and isotope ratios of noble gases, including radiogenic ⁴⁰Ar, were not measured by the Voyager remote-sensing observations. Similarly, photochemically produced trace gases in the upper atmosphere had been identified by the Voyager Infrared Radiometer and Spectrometer (IRIS)², but the fate of these constituents remained obscure. To what extent did they form more complex molecules, for example, before condensing and precipitating on the surface?

The Gas Chromatograph Mass Spectrometer (GCMS)³ on the Huygens probe was designed to help answer these and other questions concerning the atmosphere of Titan, to measure isotope abundances, and to attempt to analyse condensed phases (including isotope ratios) on the surface.

The GCMS composition and isotopic measurements provide important constraints on models of the formation of Titan and its atmosphere in particular, and on theories of the protosolar nebula and the origin and evolution of planetary systems and atmospheres in general. It is thought that planetary atmospheres are generated in two principal ways: by accretion of a portion of the solar nebula, or by impact of gas-rich planetesimals. A variation on the theme of solar nebula accretion is a subnebula in the region surrounding a giant planet such as Saturn. The giant planets seem to be an example of a

blend of solar nebula accretion and degassing from planetesimals, because Jupiter has a proportional endowment of heavy noble gases and other heavy elements (relative to hydrogen) that is greater than existed in the solar nebula. The rarity of noble gases in the atmosphere of Earth has long been viewed as strong support for a planetesimal influx, and the near absence of noble gases from Titan, as we will discuss later, provides more support for this hypothesis. Except for ³⁶Ar, heavy primordial noble gases were not detected by the GCMS instrument, yielding an upper limit for ³⁸Ar, krypton and xenon below mole fractions of 10⁻⁸. The mole fraction of ³⁶Ar is $(2.8 \pm 0.3) \times 10^{-7}$. This value will become more precise with further work.

The photochemistry of nitrogen and methane leads to the formation of complex hydrocarbons and nitriles. Methane is also key to the maintenance of the thick nitrogen atmosphere. The nitrogen atmosphere would gradually condense in the absence of warming resulting from the hydrocarbon haze and the H₂-N₂ and CH₄-N₂ collision-induced opacity in the infrared⁴. The height dependence of the methane abundance in the well-mixed atmosphere could not be determined until the Huygens probe measurements were carried out. Results of the data analysis show that the mole fraction of methane is 1.41×10^{-2} in the stratosphere, increasing below the tropopause, levelling off at 4.9×10^{-2} near the surface. The uncertainty in these methane measurements is $\pm 5\%$. Rapid increase of the methane signal after landing suggests that liquid methane exists on the surface, together with several species of higher molecular weight. GCMS isotopic measurements of carbon, nitrogen, hydrogen and argon further help to constrain atmospheric evolution and composition

¹National Aeronautics and Space Administration, Greenbelt, Maryland 20771, USA. ²University of Michigan, Ann Arbor, Michigan 48109-2143, USA. ³Institute for Meteorology and Geophysics, University of Graz, A-8010 Graz, Austria. ⁴LESIA, Observatoire de Paris-Meudon, F-92195 Meudon Cedex, France. ⁵Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85716, USA. ⁶Service d'Aéronomie du CNRS, F-91371 Verrières le Buisson Cedex, France. ⁷University of Alabama, CMC, 817 22nd Street South, Birmingham, Alabama 35205, USA. ⁸University of Hawaii, Honolulu, Hawaii 96822, USA. ⁹Laboratoire Interuniversitaire des Systèmes Atmosphériques, Université Paris 12 et Paris 7, Avenue du Général de Gaulle, F-94010 Creteil Cedex, France.

models. The isotopic ratio of $^{12}\text{C}/^{13}\text{C}$ is 82.3 ± 1 , of $^{14}\text{N}/^{15}\text{N}$ is 183 ± 5 , and of D/H is $(2.3 \pm 0.5) \times 10^{-4}$. Radiogenic ^{40}Ar was detected at a mole fraction of $(4.32 \pm 0.1) \times 10^{-5}$.

A brief description of the GCMS instrument can be found below in the Methods section. A complete description of the instrument can be found in ref. 3. Data were collected for two hours and 27 min from an altitude of 146 km to the surface. The Huygens probe and the instrument survived the surface impact, allowing data collection of gases evaporated from the surface for an additional 69 min. Here we focus on data obtained from the direct atmospheric measurements of ion source 1, which includes some results from the rare-gas cell.

Origin

The heavy primordial noble gases— $^{36,38}\text{Ar}$, Kr and Xe—have been detected and measured in meteorites, in the atmospheres of Venus and Mars, and in an over-solar abundance in Jupiter with respect to H_2 . Differing patterns of relative abundances and isotopic ratios of the gases provide insights into the origin and evolution of these objects. Hence, their measurements in the atmosphere of Titan were eagerly anticipated. Detection of these elements on Jupiter in amounts relative to C and N, essentially identical to 'solar' ratios⁵, confirmed conventional wisdom that the gases must have been present throughout the solar nebula, and should therefore have been incorporated in both Saturn and Titan. Thus, the following results based on the direct atmospheric data and the analysis of the rare-gas cell data were unexpected: the mole fraction of ^{36}Ar is $(2.8 \pm 0.3) \times 10^{-7}$ (Table 1), and no traces of ^{38}Ar , Kr or Xe were detected, with preliminary upper limits on the mole fraction of 10^{-8} . So even $^{36}\text{Ar}/^{14}\text{N}$ is about six orders of magnitude less than the solar ratio of 23. A sample mass spectrum from the stratosphere, averaged from altitudes of 130 to 120 km (~ 4.2 – 5.6 hPa ambient pressure), is shown in Fig. 1a. Major peaks at m/z of 28, 14 and 16, 15, 13, 12 show the presence of N_2 and CH_4 respectively. Figure 1b shows a spectrum from the rare-gas cell analysis, where the primordial noble gases other than a trace of ^{36}Ar and other heavy molecules are absent.

This result is especially interesting because of the huge, nitrogen-dominated atmosphere and because approximately 50% of the mass of Titan is in the form of water ice, known to be a potentially efficient carrier of noble gases^{6,7}. The low upper limits for noble gases may have implications for the origin of nitrogen on Titan, an issue that was one of the specific objectives of the exploration⁸ of Titan. Did the nitrogen arrive in planetesimals as N_2 or as a mixture of nitrogen compounds, presumably dominated by NH_3 ?

Considering the formation of the icy planetesimals that built Titan in the Saturn subnebula, we note that direct condensation of gases, trapping in amorphous ice, or formation of clathrate hydrates would all have the effect of capturing noble gases together with N_2 (refs 7, 9–12). Thus the low abundance of primordial noble gases implies that the nitrogen was captured as NH_3 and in other non- N_2 -bearing compounds. Subsequent photolysis¹³ in a hot proto-

atmosphere generated by the accreting Titan¹⁴ or possibly impact-driven chemistry¹⁵ of NH_3 led to the nitrogen atmosphere we have on Titan today.

The formation of clathrate hydrates in the feeding zones of the giant planets depends upon the efficiency of the trapping of volatiles by microscopic icy grains. It has been suggested that, based on the non-uniform enrichment in C, N and S in Saturn, the solar nebula at 10 AU provided conditions such that only CH_4 and Xe formed clathrates and were trapped in the planetesimals that formed Titan¹¹. The discovery of a small amount of ^{36}Ar in the atmosphere of Titan will allow this model to be refined. An alternative scenario suggests that icy planetesimals formed at temperatures sufficiently warm ($T > 75$ K) that noble gases and CH_4 simply were not captured, whereas NH_3 and CO_2 were. The presence of some CO_2 in the outer solar nebula is suggested by its detection in comets¹⁶.

This model, requiring *in situ* formation of CH_4 , can be tested by studies of the deuterium to hydrogen ratio D/H, because it implies that the D/H in methane should be the same as that in the water ice that formed Titan, allowing for any fractionation that may have occurred in producing the methane. In principle, the water-ice value could have been measured directly by the GCMS after impact, if the end of the inlet tube had been in direct contact with surface ice. Evidently this was not the case, because the GCMS did not detect any H_2O vaporized from the surface after probe impact. Most of the hydrogen it measured at $m/z = 2$ and 3 must have come from the fragmentation of CH_4 , but some contribution from the interior may

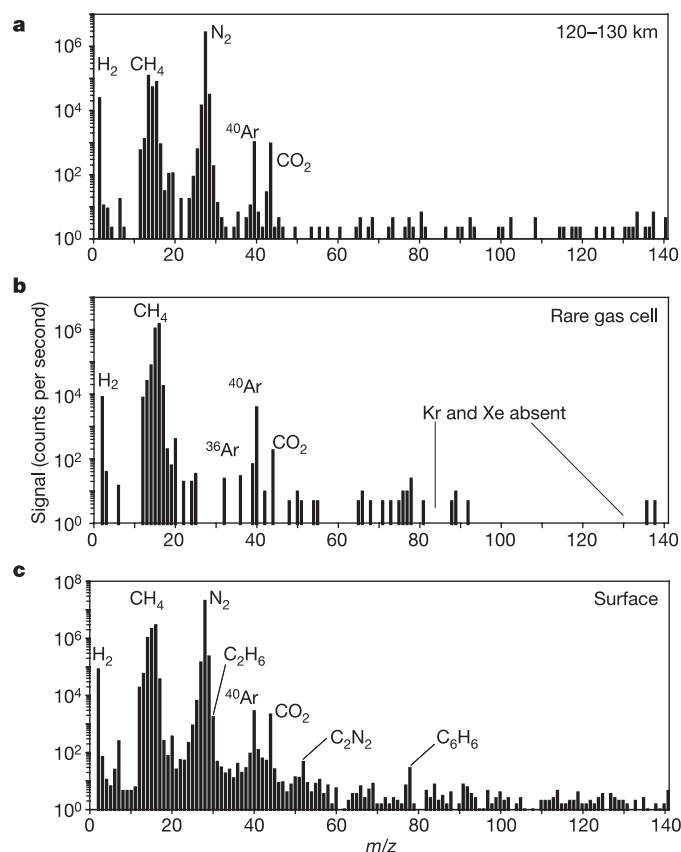


Figure 1 | Sample-averaged mass spectra, showing ion count rates per second versus mass per unit charge (m/z) from direct atmospheric sampling. a, An upper atmosphere spectrum from altitudes of approximately 120 to 130 km, averaging 104 mass spectra over 244 s. b, The rare-gas cell measurements (about 75–77 km, averaging 43 mass spectra over 81 s), showing the lack of heavy primordial noble gases. c, A surface spectrum, averaged over 70 min (432 mass spectra) from surface impact until loss of signal.

Table 1 | GCMS determination of given ratios

Ratio	GCMS	Altitude for GCMS calculations (km)	Titan/Earth
$^{14}\text{N}/^{15}\text{N}$	183 ± 5	40.9–35.9	0.67
$^{12}\text{C}/^{13}\text{C}$	82.3 ± 1	18.2–6.14	0.915
D/H	$(2.3 \pm 0.5) \times 10^{-4}$	124.9–66.8	1.44
$^{36}\text{Ar}/(\text{N}_2 + \text{CH}_4)$	$(2.8 \pm 0.3) \times 10^{-7}$	75–77 (rare-gas cell)	7.0×10^{-3}
$^{40}\text{Ar}/(\text{N}_2 + \text{CH}_4)$	$(4.32 \pm 0.1) \times 10^{-5}$	18 (to surface)	3.61×10^{-3}

For the GCMS values, the errors quoted represent one standard deviation. The altitudes for which the GCMS ratios were calculated were provided by the HASI instrument. The altitude ranges represent periods of least statistical error and best instrument conditions for those data. The Titan values are those measured by the Huygens GCMS (this paper). The $^{12}\text{C}/^{13}\text{C}$ for the Earth is the PDB standard inorganic value of 89.9.

also have been present. The proportions of these two sources will be investigated.

The measured value of $D/H = (2.3 \pm 0.5) \times 10^{-4}$ (Table 1) from HD/H_2 , is well within the uncertainties of the early ground-based $(1.6^{+1.6}_{-0.8}) \times 10^{-4}$ (ref. 17) and Voyager $(1.5^{+1.5}_{-1.0}) \times 10^{-4}$ (ref. 18) results obtained from observations of CH_3D and CH_4 in two different regions of the infrared spectrum of Titan. More recent remote-sensing measurements have led to apparently lower values: $(7.75 \pm 2.25) \times 10^{-5}$ with the Infrared Echelle Spectrometer (IRSHELL)¹⁹ and $(8.7^{+3.2}_{-1.9}) \times 10^{-5}$ with the Infrared Space Observatory/Short Wavelength Spectrometer (ISO/SWS)²⁰. The D/H on Titan is therefore an order of magnitude higher than the value in the solar nebula H_2 , slightly less than the 3.2×10^{-4} found in H_2O from Oort-cloud comets²¹, and very close to the Standard Mean Ocean Water (SMOW) terrestrial value of $D/H = 1.6 \times 10^{-4}$. Using this value of D/H with the new information on nitrogen isotopes and noble gas abundances will enable much-improved models for the origin and evolution of the atmosphere of Titan to be developed.

Evolution and composition

The evolutionary processes that led to the present Titan atmosphere depended greatly on the interaction of N_2 and CH_4 and their byproducts. Establishing an altitude profile of these constituents is critical to understanding these processes. Measurement of the isotopic ratios of the constituents, such as $^{12}C/^{13}C$ and $^{14}N/^{15}N$, is also necessary to understand the chemical processes that have shaped Titan over time. In addition, the measurement of ^{40}Ar assists in deciphering the interaction between interior planetary processes and the atmosphere.

Methane was measured with the GCMS by monitoring the mass peak at $m/z = 16$. Contributions of mass spectral fragments of any heavier molecules to this signal are negligible. The methane mole fraction was then computed from the counting rates at $m/z = 16$ and those at $m/z = 28$, after correcting for counter overflow at high count rates at $m/z = 28$ and accounting for instrument calibration factors. The result is shown as a function of altitude in Fig. 2. Measurement errors are not yet fully evaluated. Systematic deviations resulting from calibration errors are estimated to be less than $\pm 5\%$. Errors due to counting statistics are negligible.

In the stratosphere, methane was found to be uniformly mixed, with a mole fraction of 1.41×10^{-2} (Fig. 2). This finding agrees with the stratospheric CH_4 measurement by the Composite Infrared Spectrometer (CIRS) on the Cassini Orbiter $(1.6 \pm 0.5) \times 10^{-2}$ (ref. 22). At 32 km, the CH_4 mole fraction began to increase gradually, with a more rapid increase observed at approximately 16 km altitude, as indicated by the change in slope in the methane signal relative to the nitrogen signal in the Fig. 2 inset. The CH_4 mole

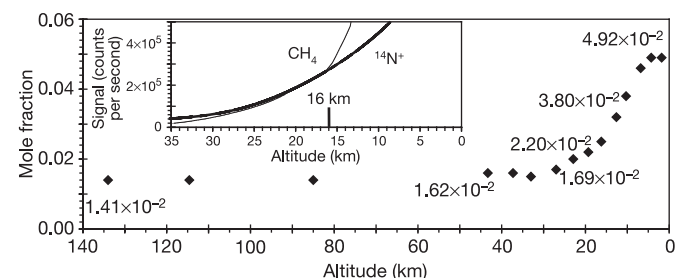


Figure 2 | The mole fraction of methane to nitrogen in the Titan atmosphere is plotted versus altitude. The CH_4 mole fraction is 1.41×10^{-2} in the stratosphere. It begins increasing below 32 km. At about 8 km, it reached a plateau of about 4.9×10^{-2} . The inset shows an increase of methane at 16 m/z , when compared to nitrogen (in this case $^{14}N^+$) at $m/z = 14$, near 16 km. This is probably due to condensates evaporating in the inlet system of the mass spectrometer as the Huygens probe passed through the methane haze.

fraction continued to rise until approximately 8 km, where it reached a value of 4.92×10^{-2} . It then remained relatively constant until the surface impact of the Huygens probe (Fig. 2). Methane was found to be subsaturated at the surface, with a relative humidity of approximately 45%. The observed behaviour of the mixing ratio—uniform up to 8 km and then declining above this altitude—appears to be a classic example in thermodynamics, indicating that methane reached its lifting condensation level, that is, near 100% relative humidity, at approximately 8 km. Indeed, the GCMS-measured CH_4 mole fraction at this level is consistent with the methane saturation values within the range of uncertainty of the Huygens Atmospheric Structure Instrument (HASI)-derived temperatures.

Therefore some haze formation is expected to take place at 8 km. The concentration of haze would probably be low due to eventual fall-out from the atmosphere, and it could escape detection by other means. The methane condensate appears to extend to an altitude of 16 km, as indicated by an increase in the gradient in the CH_4 mole fraction at 16 km. We interpret this increase as being due to the effect of evaporation of methane condensate entering the GCMS heated inlet as the probe descended through the top of the methane haze, which would slightly increase the apparent mole fraction in that region. This contribution is not easily quantitatively assessed. The liquid-binary mixture of CH_4-N_2 could potentially have a lower CH_4 vapour pressure above 'pure liquid methane'^{23,24} than that due to pure CH_4 alone. However, lack of knowledge of many factors, including the extent of pure liquid methane, concentration of N_2 dissolved in this liquid, and the presence of other impurities (many of which have already been detected by the GCMS), makes it difficult to readily ascertain the effect of the CH_4-N_2 mixture on the CH_4 vapour pressure.

The isotopic ratios of $^{12}C/^{13}C$ and $^{14}N/^{15}N$ in the atmosphere were determined by analysing the GCMS measurements of methane ($^{12}CH_4$ and $^{13}CH_4$ at m/z values of 16 and 17) and molecular nitrogen ($^{14}N^{14}N$ and $^{14}N^{15}N$ at m/z values of 28 and 29). In addition, radiogenic ^{40}Ar was detected at $m/z = 40$. The isotopic ratio of $^{12}C/^{13}C$ is 82.3 ± 1 , and of $^{14}N/^{15}N$ is 183 ± 5 . Radiogenic ^{40}Ar was detected at a mole fraction of $(4.32 \pm 0.1) \times 10^{-5}$ (Table 1). The errors listed are derived from signal statistics (one standard deviation), which is the largest source of error. A detailed error analysis including systematic errors is in progress.

Ground-based observations of Titan have revealed a huge fractionation (by a factor of about four) of the nitrogen isotopes in HCN, suggesting the escape of a massive early atmosphere^{25,26}. In contrast, for reasons that are not yet clear, $^{12}C/^{13}C$ in HCN was measured from Earth to lie in the range 70 to 120, embracing the terrestrial ratio²⁷. The GCMS did not measure isotope ratios in HCN on Titan, but it was found from the GCMS measurements that in N_2 the depletion of ^{14}N is much less, leading to a value of $^{14}N/^{15}N$ only 1.5 times less than the terrestrial value. Apparently photochemistry is strongly enriching the heavy isotope of nitrogen in HCN. As the terrestrial value appears to represent the value in nitrogen compounds, especially NH_3 , in the solar nebula²⁸, it seems the right standard to use for comparison with Titan.

With this assumption, calculations using the ground-based HCN-derived $^{14}N/^{15}N$ ratio of four times less than the terrestrial value predicted a primitive atmosphere between 20 and 100 times more massive than today²⁶. Using 1.5 (from the GCMS-measured value in the major nitrogen reservoir, N_2) instead of four leads to an estimate for the mass of the primitive atmosphere of between two and ten times today's value. Therefore, it is estimated that perhaps several times the present mass of the atmosphere was lost over geologic time²⁶. The ratio of $C^{18}O/C^{16}O$ is approximately two times greater than the terrestrial value as inferred from ground-based millimetre measurements²⁹. Modelling has been carried out of the evolution of heavy isotopic isomers of CO (ref. 30), with the constraint to obtain a telluric $^{12}C/^{13}C$ today, and to reproduce the observed $C^{18}O/C^{16}O$ ratio. The conclusion was that the depletion by a factor of 2 ± 0.5 in

$^{18}\text{O}/^{16}\text{O}$ was roughly consistent with a depletion by a factor of 1.5 in $^{15}\text{N}/^{14}\text{N}$, but clearly more work on this problem is needed. In contrast, the GCMS-measured value for $^{12}\text{C}/^{13}\text{C}$ is only slightly smaller than the terrestrial value.

Because photochemistry destroys methane irreversibly on Titan³¹, so that its lifetime in the atmosphere is only 10–100 Myr (refs 32, 33), and because the carbon in CH_4 does not show the same kind of isotopic fractionation as the nitrogen and oxygen isotopes do, methane must be continually or periodically replenished on Titan. Models for atmospheric escape on Titan need to address the cause of the slight depletion of ^{12}C .

The value of $^{12}\text{C}/^{13}\text{C}$ in methane provides no support for suggestions of an active biota on Titan. It takes less energy to form a chemical bond between two ^{12}C atoms than between ^{12}C and ^{13}C , so complex organic molecules associated with biological processes on earth show an enrichment in ^{12}C —the $^{12}\text{C}/^{13}\text{C}$ ratio is greater than the Pee Dee Belemnite (PDB) inorganic standard value of 89.9, and could be as high as 95. Therefore the assumption that such enrichment will occur in carbon-based non-terrestrial biology seems reasonable.

We do not find this enrichment in the methane of Titan. Instead, a geological source for methane, with a possible clathrate reservoir as storage in the interior of Titan is favoured. As occurs on Earth^{34–36}, serpentinization that releases hydrogen from water while oxidizing iron- or magnesium-bearing minerals could produce methane through a Fischer–Tropsch reaction of the H_2 with CO_2 (ref. 37), or reduction of carbon grains in the crustal rocks in the interior.

Alternatively, methane may have been captured from the subnebula in the form of clathrate hydrates^{25,38,39} that now float on a plausible subcrustal ammonia–water ocean (G. Tobie, personal communication)⁹. If the total carbon inventory of Titan exhibits the same ratio to nitrogen of about 20 as is found on Earth, Venus and Halley's comet^{10,40}, and approximately five times the present atmospheric N has been lost, and the total aerosol deposit produced in the last 4.5 billion years is $8\text{--}80\text{ kg cm}^{-2}$ (D. F. Strobel, personal communication), then only 0.8–6% of the total carbon budget has passed through the atmosphere in the last 4.5 billion years, ultimately ending up as solids or liquids on the surface of Titan. Hence, in the clathrate model, we anticipate a layer of several kilometres of methane clathrates still present on top of a plausible subcrustal ocean.

Radiogenic argon, ^{40}Ar , was detected by the GCMS below 18 km (Table 1). Above that altitude the measurement was obscured by ^{40}Ar instrument background. Radiogenic ^{40}Ar is a decay product of ^{40}K , which has a half-life of 1.28 billion years. Thus, most of the radiogenic argon on Titan has been produced over the lifetime of the Solar System and is potentially an indicator of the extent to which outgassing of volatile elements has occurred from the deep interior, where the rock (hence the potassium) should reside.

If the rocky component of the interior of Titan has the same composition as that of the Earth and has outgassed to the same extent, ^{40}Ar should be about ten times more abundant than

measured, comprising approximately 0.05% of the atmosphere⁴¹ (corrected for loss of nitrogen). If the interior was warm enough in the past for a liquid water or liquid water–ammonia mantle to have reached all the way to the rocky core, potassium could have leached into the liquid and the radiogenic fraction expressed as ^{40}Ar outgassed to the surface⁴². The ^{40}Ar signature from the GCMS data may thus reflect a complex multi-step process that brought first the parent element, then the argon itself, upward through liquid and solid water layers to the surface. But just the presence of the ^{40}Ar at the levels seen is a strong indication of the geological activity of Titan, and is consistent with the requirement to replace the atmospheric methane episodically as described above. The apparent evidence for cryo- (water or water–ammonia) volcanism seen in the Cassini orbiter radar images⁴³ and the Visual and Infrared Mapping Spectrometer (VIMS) observations⁴⁴ provides one possible process for release of both gases from the interior.

Surface

Although numerous heavy hydrocarbons and nitriles are expected to be produced photochemically above 500 km altitude^{33,45}, and have indeed been detected at infrared wavelengths^{18,22} and in the Cassini Ion and Neutral Mass Spectrometer (INMS) fly-bys⁴⁶, very few avoid condensation at the low temperatures of the lower stratosphere (below ~ 200 km) and the tropopause^{33,47}. Therefore it is not surprising that the GCMS measurements, which were done below 146 km, do not show the presence of many heavy molecules (Fig. 1a). On the other hand, in the measurements at the surface (Fig. 1c), there is a greater likelihood of detecting some of the heavier hydrocarbons that have precipitated out of the atmosphere.

Upon the impact of the Huygens probe, the inlet line heater also heated the areas near the inlet port, including the Titan surface in the near vicinity of the GCMS inlet, either by contact or radiation. The exact nature of the thermal contact of the GCMS inlet port with the surface material is not known, and so nor is the actual contact temperature to which the surface material was heated. The inlet line temperature at the physical position of the heater, which was monitored, stabilized at 80°C . The actual temperature distribution of the inside surfaces along the inlet line was not measured. Prior to the thermal equilibration of the inlet line, the methane count rate increased by 40% while the nitrogen count rate remained constant. This increased value for methane remained nearly constant for about 50 min, and then gradually decreased to about 25% above the pre-impact value at 69 min after impact near the end of the data transmission. The methane and nitrogen signal versus time near impact is shown in Fig. 3. Although the thermal environment at or near the inlet port of the GCMS on the surface cannot be known, we suggest that the behaviour of methane in the surface measurements indicates the presence of liquid methane mixed with the surface material. The slight decrease in CH_4 after 50 min on the surface (not shown in Fig. 3) may be due to the depletion of liquid methane in the immediate vicinity of the GCMS inlet, for example, in a wet, loosely aggregated aerosol deposit. Evaporation of methane previously condensed or trapped inside the sampling system during the descent instead of evaporation from the surface is unlikely because of the high temperature inside the sample line and the long steady rate of evaporation that was observed.

The mass spectra taken on the surface (Fig. 1c) show mass peaks characteristic of more complex molecules. Ethane was firmly identified. Cyanogen (C_2N_2), benzene, and carbon dioxide have been tentatively identified, and work is continuing to identify other constituents. Each of these is much less volatile than methane and hence would be expected to have a smaller signature, regardless of their bulk abundance on the surface.

Future work

There is a large body of data from the GCMS that remains to be analysed in detail, including the gas chromatograph data, enrichment

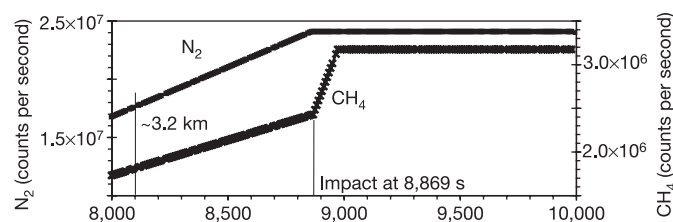


Figure 3 | Pulse count rates of nitrogen and methane are shown versus time. The methane count rate increases 2 min after impact by about 40%. This correlates with the rapid increase in temperature of the GCMS heated inlet, which heated the surface in the vicinity of the inlet port.

cell data, and surface data. For example, heavy hydrocarbons with mole fractions less than 100 p.p.b. may yet be identified in the enrichment cell data. Collaborations with the other probe instruments will continue to provide new and refined models of the Titan atmosphere, and will improve our understanding of the history and present chemistry of both atmospheric and surface processes.

METHODS

The GCMS uses a quadrupole mass filter with a secondary electron multiplier detection system and a gas sampling system providing continuous direct atmospheric composition measurements and batch sampling through three gas chromatographic columns³. The mass spectrometer used five electron impact ion sources with available electron energies of either 70 or 25 eV. Three ion sources served as detectors for the gas chromatographic columns and two were dedicated to direct atmosphere sampling and Aerosol Collector Pyrolyser (ACP)⁴⁸ gas sampling, respectively. The multiple ion source approach allowed rapid switching between sample systems and prevented cross-contamination. The instrument was also equipped with a chemical scrubber cell for noble-gas analysis and a sample-enrichment cell for selective measurement of high-boiling-point carbon-containing constituents. The mass filter produced flat-top mass peaks that allowed rapid scanning in 5-ms steps of unit values of mass to charge (m/z) ratios over a range from 2 to 141. The nominal detection threshold was at a mixing ratio of 10^{-8} .

Pressure reduction from the ambient pressure, ~ 3 to $\sim 1,500$ hPa (~ 1.5 bar), during the probe's descent to the vacuum level of $<10^{-4}$ hPa was achieved with micrometre-sized glass capillary arrays. A choice of two capillary arrays of different gas conductance was used for the direct atmosphere ion source to cover the wide pressure range during the descent. Gases were removed from the ion sources by conductance limited getter and sputter ion pumps. The maximum ion source operating pressure was 10^{-4} hPa and the mass filter pressure was always kept below 10^{-6} hPa. The ambient atmosphere was sampled from flow through a tube whose inlet was near the apex of the probe fairing and whose outlet was at the rear of the probe. The pressure difference created between the inlet and outlet owing to the motion of the Probe caused the atmospheric gas to flow through the tube during the descent. To prevent condensation and to cause rapid evaporation of condensates that might flow through the gas sampling system, the inlet section, upstream from the sampling area, was heated up to 80 °C, and reached temperatures that depended on gas flow rates through the inlet line.

The measurement sequence was pre-programmed. Direct atmospheric samples were taken nearly continuously during the entire descent, interrupted only when the ACP samples and the contents of the rare-gas and the sample-enrichment cells were analysed.

The sample inlet system and the mass spectrometer were sealed under vacuum until exposed to the ambient atmosphere after jettison of the probe's heat shield. The descent sequence was properly executed during the mission. However, ion source 5, serving as the detector for the N_2 -CO separation column, ceased operation owing to an electrical malfunction early in the descent. This resulted in the loss of all data from this column, and in particular the measurement of the CO height profile. Coincidentally, external perturbations affecting the Huygens probe motion were also experienced at the same time.

Received 2 June; accepted 3 August 2005.

Published online 30 November 2005.

- Hunten, D. M., et al. in *Saturn* (eds Gehrels, T. & Shapley Matthews, M.) 671–759 (Univ. Arizona Press, Tucson, 1984).
- Kunde, V. G. et al. C_4H_2 , HC_3N and C_2N_2 in Titan's atmosphere. *Nature* **292**, 686–688 (1981).
- Niemann, H. B. et al. The Gas Chromatograph Mass Spectrometer for the Huygens probe. *Space Sci. Rev.* **104**(1), 553–591 (2002).
- Lorenz, R. D., McKay, C. P. & Lunine, J. L. Photochemically driven collapse of Titan's atmosphere. *Science* **275**, 642–644 (1997).
- Niemann, H. B. et al. Chemical composition measurements of the atmosphere of Jupiter with the Galileo Probe Mass Spectrometer. *Adv. Space Res.* **21**, 1455–1461 (1998).
- Lunine, J. I. & Stevenson, D. J. Clathrate and ammonia hydrates at high pressure: Application to the origin of methane on Titan. *Icarus* **70**, 61–77 (1987).
- Bar-Nun, A., Kleinfeld, A. I. & Kochavi, E. Trapping of gas mixtures by amorphous ice. *Phys. Rev. B* **38**, 7749–7754 (1988).
- Owen, T. & Gautier, D. Touring the Saturn system: The atmospheres of Saturn and Titan. *Space Sci. Rev.* **104**, 347–376 (2002).
- Tobie, G., Grasset, O., Lunine, J. I., Mocquet, A. & Sotin, C. Titan's internal structure inferred from a coupled thermal-orbital model. *Icarus* **175**, 496–502 (2005).
- Owen, T. & Bar Nun, A. Comets, impacts and atmospheres. *Icarus* **116**, 215–226 (1995).
- Hersant, F., Gautier, D. & Lunine, J. Enrichment in volatiles in the giant planets of the Solar System. *Planet. Space Sci.* **52**, 623–641 (2004).
- Mousis, O., Gautier, D. & Bockelée-Morvan, D. Turbulent model of the Saturn subnebula: Implications for the origin of methane in Titan's atmosphere. *Icarus* **156**, 162–175 (2002).
- Atreya, S. K., Donahue, T. M. & Kuhn, W. R. Evolution of a nitrogen atmosphere on Titan. *Science* **201**, 611–613 (1978).
- Kuramoto, K. & Matsui, T. Formation of a hot proto-atmosphere on the accreting giant icy satellite: Implications for the origin and evolution of Titan, Ganymede, and Callisto. *J. Geophys. Res.* **99**(E10), 21183–21200 (1994).
- McKay, C. P., Scattergood, T. W., Pollack, J. B., Borucki, W. J. & Van Ghysseghem, H. T. High temperature shock formation of N_2 and organics on primordial Titan. *Nature* **332**, 520–522 (1988).
- Bockelée-Morvan, D., Crovisier, J., Mumma, M. J. & Weaver, H. A. The composition of cometary volatiles in *Comets II* (eds Festou, M., Weaver, H. A. & Keller, H. U.) 391–423 (Univ. Arizona Press, Tucson, 2004).
- de Bergh, C., Lutz, B., Owen, T. & Chauville, J. Monodeuterated methane in the outer solar system. III. Its abundance on Titan. *Astrophys. J.* **311**, 501–510 (1986).
- Coustenis, A., Bézard, B. & Gautier, D. Titan's atmosphere from Voyager infrared observations: II. The CH_3D abundance and D/H ratio from the 900–1,200 cm^{-1} spectral region. *Icarus* **82**, 67–80 (1989).
- Orton, G. et al. *The D/H ratio on Titan from ISO and IRSHELL Data* 81 (ESA report SP 338, Noordwijk, 1992).
- Coustenis, A. et al. Titan's atmosphere from ISO mid-infrared spectroscopy. *Icarus* **161**, 383–403 (2003).
- Meier, R. & Owen, T. Cometary deuterium. *Space Sci. Rev.* **90**, 33–43 (1999).
- Flasar, F. M. et al. Titan's atmospheric temperatures, winds, and composition. *Science* **308**, 975–978 (2005).
- Kouvaris, L. C. & Flasar, F. M. Phase equilibrium of methane and nitrogen at low temperatures—Application to Titan. *Icarus* **91**, 112–124 (1991).
- Sagan, C., Thompson, W. R. & Khare, B. N. Titan: a laboratory for prebiological organic chemistry. *Acc. Chem. Res.* **25**, 286–292 (1992).
- Marten, A., Hidayat, T., Biraud, Y. & Moreno, R. New millimeter heterodyne observations of Titan: Vertical distributions of nitriles HCN , HC_3N , CH_3CN , and the isotopic ratio $^{15}N/^{14}N$ in its atmosphere. *Icarus* **158**, 532–544 (2002).
- Lunine, J. I., Yung, Y. L. & Lorenz, R. D. On the volatile inventory of Titan from isotopic abundances in nitrogen and methane. *Planet. Space Sci.* **47**, 1291–1303 (1999).
- Hidayat, T. et al. Millimeter and submillimeter heterodyne observations of Titan: Retrieval of the vertical profile of HCN and the $^{12}C/^{13}C$ ratio. *Icarus* **126**, 170–182 (1997).
- Owen, T., Mahaffy, P. R., Niemann, H. B., Atreya, S. K. & Wong, M. Protosolar nitrogen. *Astrophys. J.* **553**, L77–L79 (2001).
- Owen, T., Biver, N., Marten, A., Matthews, H. & Meier, R. *Saturn VI (Titan)*. (ed. Green, D. W. E.) *IAU Circ.* **7306**, 3, (1999).
- Wong, A.-S., Morgan, C. G., Yung, Y. L. & Owen, T. Evolution of CO on Titan. *Icarus* **155**, 382–392 (2002).
- Strobel, D. F. The photochemistry in the atmosphere of Titan. *Icarus* **21**, 466–470 (1974).
- Yung, Y. L., Allen, M. & Pinto, J. P. Photochemistry of the atmosphere of Titan—Comparison between model and observations. *Astrophys. J. Suppl. Ser.* **55**, 465–506 (1984).
- Wilson, E. H. & Atreya, S. K. Current state of modelling the photochemistry of Titan's mutually dependent atmosphere and ionosphere. *J. Geophys. Res.* **109**, E06002, doi:10.1029/2003JE002181 (2004).
- Boetius, A. Lost city life. *Science* **307**, 1420 (2005).
- Kelley, D. S. et al. A serpentinite-hosted ecosystem: the lost city hydrothermal field. *Science* **307**, 1428–1434 (2005).
- Chiodini, G. Carbon dioxide Earth degassing and seismogenesis in central and southern Italy. *Geophys. Res. Lett.* **31**, L07615 (2004).
- Foustoukos, D., Seyfried, W. E. & Dionysis, I. Hydrocarbons in hydrothermal vent fluids: the role of chromium-bearing catalysts. *Science* **304**, 1002–1005 (2004).
- Hersant, F. G., Gautier, D. & Lunine, J. Enrichment in volatiles in the giant planets in the Solar System. *Planet. Space Sci.* **52**, 623–641 (2004).
- Mousis, O., Gautier, D. & Bockelée-Morvan, D. An evolutionary turbulent model of Saturn's subnebula: implications for the origin of the atmosphere of Titan. *Icarus* **156**, 162–175 (2002).
- Holland, H. *The Chemical Evolution of the Atmosphere and Oceans* 39 (Princeton Univ. Press, Princeton, 1984).
- Owen, T. C. The composition and origin of Titan's atmosphere. *Planet. Space Sci.* **30**, 833–838 (1982).
- Engel, S., Lunine, J. I. & Norton, D. L. Silicate interactions with ammonia-water fluids on early Titan. *J. Geophys. Res.* **99**, 3745–3752 (1994).
- Elachi, C. et al. Cassini RADAR's first views of the surface of Titan. *Science* **308**, 970–974 (2005).
- Sotin, C. et al. Release of volatiles from a possible cryovolcano from near-infrared imaging of Titan. *Nature* **435**, 786–789 (2005).

45. Lebonnois, S. *et al.* Seasonal variations of Titan's atmospheric composition. *Icarus* **152**, 384–406 (2001).
46. Waite, J. H. *et al.* Ion neutral mass spectrometer results from the first flyby of Titan. *Science* **308**, 982–986 (2005).
47. Wilson, E. H. & Atreya, S. K. Chemical sources of haze formation in Titan's atmosphere. *Planet. Space Sci.* **51**, 1017–1033 (2003).
48. Israel, G. *et al.* Huygens Probe Aerosol Collector Pyrolyser Experiment. *Space Sci. Rev.* **104**(1), 433–468 (2002).

Acknowledgements This paper is dedicated to the memory of T. Donahue, who contributed to the planning and development of the GCMS, and died before the Huygens probe encountered Titan. We acknowledge the HASI team, who provided the atmospheric pressure–temperature–altitude data to the GCMS team. We thank F. M. Flasar, R. H. Brown and C. Sotin for providing preprints of their Cassini papers. We also thank G. Tobie for information on the story of

clathrate hydrates within Titan, F. Hersant for discussions on enrichments by clathration in giant planets and D. Strobel for his discussions of atmospheric loss. The contributions of personnel at the NASA Goddard Space Flight Center, the University of Michigan, the Ohio State University and the University of Paris are acknowledged. We thank the personnel at the European Space Research and Technology Centre (ESTEC) and the European Space Operations Centre (ESOC) for their technical support and guidance during this mission. We acknowledge NASA, ESA and CNES for support of the mission.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.B.N. (Hasso.B.Niemann@nasa.gov).

In situ measurements of the physical characteristics of Titan's environment

M. Fulchignoni^{1,2}, F. Ferri³, F. Angrilli³, A. J. Ball⁴, A. Bar-Nun⁵, M. A. Barucci¹, C. Bettanini³, G. Bianchini³, W. Borucki⁶, G. Colombatti³, M. Coradini⁷, A. Coustenis¹, S. Debei³, P. Falkner⁸, G. Fanti³, E. Flamini⁹, V. Gaborit¹, R. Grard⁸, M. Hamelin^{10,11}, A. M. Harri¹², B. Hathi⁴, I. Jernej¹³, M. R. Leese⁴, A. Lehto¹², P. F. Lion Stoppato³, J. J. López-Moreno¹⁴, T. Mäkinen¹², J. A. M. McDonnell⁴, C. P. McKay⁶, G. Molina-Cuberos¹⁵, F. M. Neubauer¹⁶, V. Pirronello¹⁷, R. Rodrigo¹⁴, B. Saggin¹⁸, K. Schwingenschuh¹³, A. Seiff[‡], F. Simões¹⁰, H. Svedhem⁸, T. Tokano¹⁶, M. C. Towner⁴, R. Trautner⁸, P. Withers^{4,19} & J. C. Zarnecki⁴

On the basis of previous ground-based and fly-by information, we knew that Titan's atmosphere was mainly nitrogen, with some methane, but its temperature and pressure profiles were poorly constrained because of uncertainties in the detailed composition. The extent of atmospheric electricity ('lightning') was also hitherto unknown. Here we report the temperature and density profiles, as determined by the Huygens Atmospheric Structure Instrument (HASI), from an altitude of 1,400 km down to the surface. In the upper part of the atmosphere, the temperature and density were both higher than expected. There is a lower ionospheric layer between 140 km and 40 km, with electrical conductivity peaking near 60 km. We may also have seen the signature of lightning. At the surface, the temperature was 93.65 ± 0.25 K, and the pressure was $1,467 \pm 1$ hPa.

Earlier Voyager fly-bys of Titan and telescopic observations indicated that Titan's atmosphere is composed of N₂ with small amounts of CH₄. The surface pressure was determined to be approximately 1,400 hPa, with a surface temperature of about 95 K decreasing to a temperature minimum of about 70 K at 40 km altitude before increasing again to about 170 K in the stratosphere^{1–3}. The atmospheric structure at high elevations (1,000–1,500 km) was inferred from the solar occultation measurements by the Voyager ultraviolet spectrometer (UVS)⁴. The middle atmosphere (200–600 km) was not well determined, although telescopic observations indicated a complex vertical structure^{5–10} and models have been used to predict the atmospheric structure in this region^{11–13}. Very little was known about the surface of Titan because it is hidden by a thick haze and is almost undetectable, except by radar sounding¹⁴ and a few infrared windows that have been observed from telescopes^{15,16}. Initial speculation was that the surface was covered by a deep hydrocarbon ocean, but infrared and radar measurements showed definite albedo contrasts—possibly consistent with lakes, but not with a global ocean. Recently, measurements by the Cassini orbiter in the near-infrared and at radar frequencies provided new results on the nature of the surface of the satellite^{17–19}.

Earlier observations showed that the surface pressure on Titan was comparable to that on the Earth, and that CH₄ formed a plausible counterpart to terrestrial H₂O for cloud and rain formation. There was also speculation on the possibility of lightning occurring in

Titan's atmosphere^{20–22} which could affect the chemical composition of the atmosphere.

In this Article, we report results from the HASI instrument on the Huygens probe²³. By monitoring the probe deceleration, the HASI instrument directly determined the density of the upper atmosphere and derived the temperature from the density scale height. In the lower atmosphere and on the surface of Titan, the HASI instrument directly measured the pressure and temperature. During the probe descent, electrical activity was monitored to search for evidence of lightning activity. A search for acoustic signals produced by any thunder or other shock waves was also conducted. A comprehensive description of the HASI instrument can be found in ref. 24.

In the upper atmosphere, the density profile is used to infer the temperature profile. Above 500 km, the temperature structure shows strong wave-like variations of 10–20 K about a mean of about 170 K. Below 500 km, the temperature increases to a relative maximum of 186 K and then reaches an absolute minimum of 70 K at 44 km. Below about 200 km, the temperature and pressure profile measured by HASI agrees with the results of the Voyager radio occultation data². The surface temperature is determined to be 93.65 ± 0.25 K, and the surface pressure is $1,467 \pm 1$ hPa. The values are within the range allowed by the uncertainties in the Voyager data¹³ owing to previous uncertainties in the mixing ratio of CH₄ and argon. Electrical conductivity measurements indicate the presence of charged particles

¹LESIA, Observatoire de Paris, 5 Place Janssen, 92195 Meudon, France. ²Université Denis Diderot - Paris 7, UFR de Physique, 2 Place Jussieu, 75006 Paris, France. ³CISAS "G. Colombo", Università di Padova, Via Venezia 15, 35131 Padova, Italy. ⁴PSSRI, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK. ⁵Department of Geophysics and Planetary Sciences, University of Tel Aviv, 69978 Tel Aviv, Israel. ⁶NASA/AMES Research Center, MS 244-30, Moffett Field, California 94035, USA. ⁷ESA Headquarters, Science Directorate, 8-10 rue Mario-Nikis, 75015 Paris, France. ⁸ESA-ESTEC, European Space Agency, Keplerlaan 1, 2200 AG Noordwijk, The Netherlands. ⁹Agenzia Spaziale Italiana, Viale Liegi 26, 00198 Roma, Italy. ¹⁰CETP-IPSL, 4 Avenue de Neptune, 94107 Saint Maur, France. ¹¹LPCE-CNRS, 3A, Avenue de la Recherche Scientifique, 45071 Orléans cedex 2, France. ¹²Finnish Meteorological Institute (FMI), Vuorikatu 15 A 00100 Helsinki, Finland. ¹³Space Research Institute, Austrian Academy of Sciences (IWF), Schmiedlstrasse 6, 8042 Graz, Austria. ¹⁴Instituto de Astrofísica de Andalucía (IAA-CSIC), PO Box 3004, 18080 Granada, Spain. ¹⁵Applied Electromagnetic Group, Department of Physics, University of Murcia, Murcia 30100, Spain. ¹⁶Institut für Geophysik und Meteorologie, Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln, Germany. ¹⁷DMFCl, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy. ¹⁸Politecnico di Milano, Dipartimento di Meccanica, Piazza Leonardo da Vinci 32, 20133 Milano, Italy. ¹⁹Center for Space Physics, Boston University, 725 Commonwealth Avenue, Boston, Massachusetts 02215, USA. [‡]Deceased.

species in an ionized layer, presumably induced by cosmic rays, and the detection of some electrical discharges.

Atmosphere

We inferred the atmospheric structure of Titan on the basis of measurements taken during entry phase and while the probe was descending under the parachutes. The atmosphere was first detected at an altitude of $\sim 1,500$ km, when it exceeded the sensitivity threshold of the accelerometer²⁵. Broadly speaking, the temperature and density of the upper atmosphere exceeded predictions. Titan's atmosphere is apparently highly stratified. The density of the upper atmosphere was derived from the probe deceleration due to aerodynamic drag force, following a method^{24,25} previously used for other planetary atmospheres such as Venus, Mars and Jupiter. The velocity as a function of time was determined by integrating the measured probe deceleration. Altitude was determined by integrating the vertical component of the velocity using the state vector of the probe provided by the Cassini navigation team. The entry altitude has a 1σ uncertainty of about 30 km; we adjusted the nominal entry altitude within this standard deviation to ensure consistency between the entry phase and descent phase measurements²⁶.

The derived density profile is shown in Fig. 1, with a comparison of the engineering model¹³ obtained from the reanalysis^{2-4,11,12} of Voyager data (radio occultations, infrared interferometry (IRIS) and UVS spectrometers). In the upper part of the atmosphere down to an altitude of about 500 km, the HASI measurements show density values systematically higher than those expected. Pressures were obtained from the density profile under the assumption of hydrostatic equilibrium and the knowledge of planetary gravity (1.354 m s^{-2} at surface level), mass ($1.35 \times 10^{23} \text{ kg}$) and radius (2,575 km). Temperatures were derived from the pressures, the inferred densities and the equation of state of a perfect gas using the atmospheric mean molecular weight, as a function of altitude given by the engineering model. The pressure versus temperature profile of Titan's atmosphere is shown in Fig. 2. The thermosphere is characterized by the presence of temperature variations due to inversion layers or other dynamic phenomena (such as gravity waves and gravitational tides) between 500 km and 1,020 km.

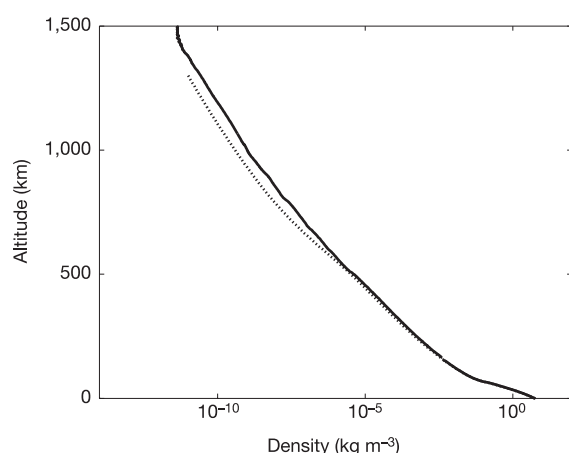


Figure 1 | The atmospheric density profile of Titan as measured by HASI. The density profile as derived from HASI measurements (solid line) is shown in comparison with the engineering model of Titan's atmosphere¹³ derived from Voyager 1 data^{2-4,11,12} (dashed line). Density in the upper part of the atmosphere is derived from the ACC accelerometer data. The threshold density was $5 \times 10^{-12} \text{ kg m}^{-3}$. The uncertainty on the density determination²⁵ is of the order of 10%, mainly due to the uncertainty on the aerodynamic drag coefficient and on the probe velocity. Density values relevant to the lower atmosphere, below 160 km, have been inferred from HASI direct measurements of pressure and temperature with the assumption of hydrostatic equilibrium and real gas law²⁹.

Temperatures in this region are generally higher than those predicted by the engineering model, with a minimum value of 152 K at an altitude of ~ 490 km (2×10^{-3} hPa, which could mark the mesopause) and then increase down to the stratopause (~ 186 K at 250 km, 0.3 hPa). In the region between the lower part of the mesosphere and the upper part of the stratosphere, the temperatures are 5–10 K higher than those predicted by the model¹².

The temperature gradient profile, shown in Fig. 3, exhibits in general a cut-off at the dry adiabatic lapse rate, implying that fluctuations lead to marginally convective instabilities. The inversion layers in the upper atmosphere are clearly visible, with strong peaks towards positive values. The peak at 510 km corresponds to the inversion layer already observed from the ground on 14 November 2003 when Titan occulted two bright Tycho stars¹⁰. These lines of evidence all indicate that Titan's atmosphere is highly stratified.

After the parachute deployment and heatshield separation, the temperature sensors²⁷ and pressure sensors²⁸ were directly exposed to Titan's environment during the entire descent under parachute. The altitude and velocity are derived from these measurements, the hypothesis of hydrostatic equilibrium, and the equation of state for a real gas²⁹, given the atmospheric mean molecular weight measured by the Gas Chromatograph-Mass Spectrometer (GCMS)³⁰. The measured pressure and temperature profiles shown in Figs 4 and 5 connect well with the profiles derived during the entry phase. From the surface up to about 150 km altitude, the HASI temperatures are in very good agreement (within 1–2 K) with the temperature measurements obtained by the Voyager 1 radio occultation assuming a pure nitrogen atmosphere².

The temperature minimum of 70.43 ± 0.25 K is reached at the tropopause (~ 44 km, 115 hPa). Figure 6 shows the temperature lapse rate in the low atmosphere. A number of inversion layers in the lower stratosphere and the strong increase in temperature with altitude between 80 km and 60 km are visible. Below 200 km, the fine

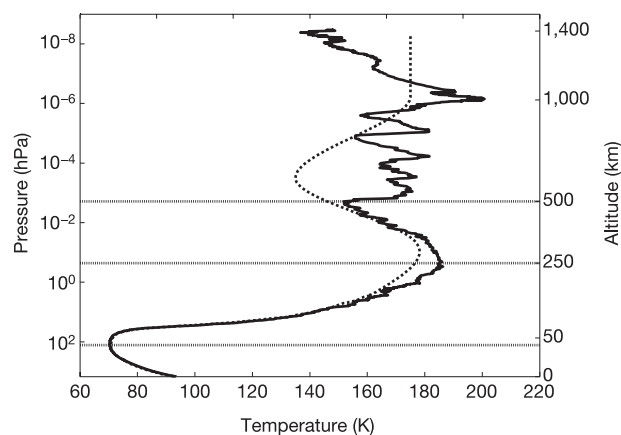


Figure 2 | The atmospheric temperature profile. The temperature profile as measured by HASI (solid line) is shown compared to Titan's atmospheric engineering model¹³ (dashed line). In the upper atmosphere (above 160 km), temperature and pressure have been derived from the density using the ideal gas equation; below 160 km, temperature data are direct measurements collected by the TEM sensor. The temperature profile in the upper atmosphere (thermosphere) is characterized by several temperature variations due to inversion layers and other dynamic phenomena (for example, gravity waves and tides). Temperatures in this region are higher than those predicted by the model. The virtual absence of a mesosphere (in contrast with the theoretical models' predictions^{11,12}) and the wave-like nature of the temperature profile suggest that the region in Titan's atmosphere above 250 km may not be dominated by radiative processes and may be strongly influenced by wave activity. Thus the structure that we observe may vary with time. The horizontal lines mark the mesopause (152 K at 490 km), the stratopause (186 K at 250 km) and the tropopause (70.43 K at 44 km).

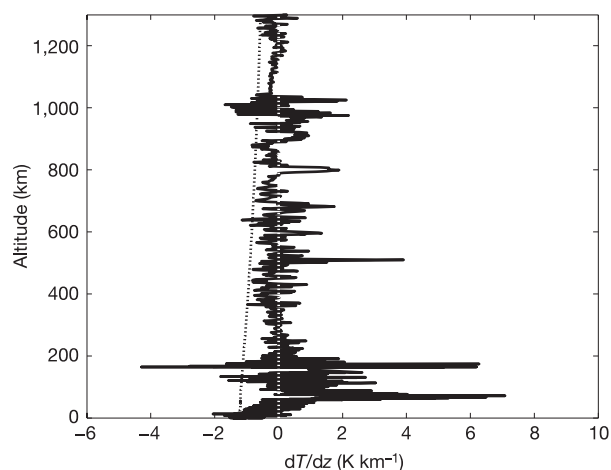


Figure 3 | The temperature lapse rate. The temperature gradient dT/dz was computed from the HASI temperature ($T(z)$) profile, and the altitude (z) was derived from the Huygens trajectory reconstruction²⁶. The spatial resolution of the HASI measurements is of the order of 20 km from the top of the atmosphere down to the 400-km altitude level, decreasing down to 1 km at the 160-km level²⁴. The profile shows in general a cut-off at the dry adiabatic lapse rate (dotted line), implying that fluctuations may lead to convective instabilities. The line at zero temperature variation is shown in white against the black curve. Six inversion layers in the upper atmosphere (at about 510, 600, 680, 800, 980 and 1,020 km) could be detected by strong peaks towards positive values. The strong lower inversion layer (4 K km^{-1} at ~ 510 km) corresponds to the feature already observed from the ground during Titan's stellar occultations¹⁰. The strong peaks between the 160- and 110-km levels correspond to the parachute deployment sequence.

structure seen in Fig. 5 provides evidence for a regime of gravity waves similar to those observed in the Voyager radio occultation data^{31,32}. Turbulence due to shear instability (Kelvin–Helmholtz instability) is expected wherever the vertical shear of the wind speed is large. The wind shear measured by the Doppler Wind Experiment³³ is sufficiently large that the features present between 50 and 150 km are likely to be related to turbulence.

The vertical resolution of the temperature measurement was sufficient to resolve the instantaneous structure of the planetary boundary layer. On the basis of the nearly constant values of the

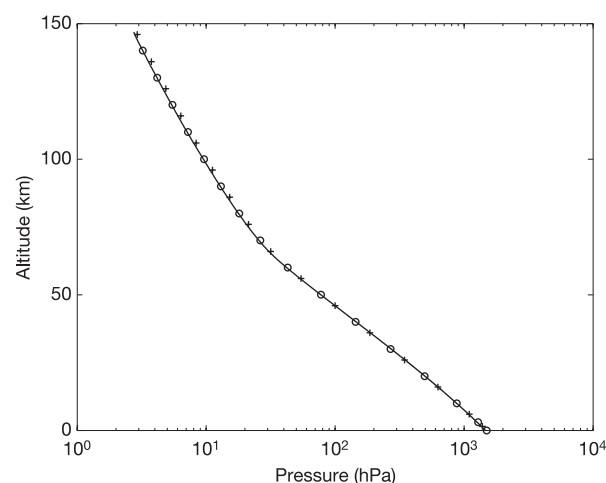


Figure 4 | Pressure profile of the lower atmosphere as measured by the Pressure Profile Instrument (PPI)²⁸. Measurements (solid line) corrected for dynamic effects are shown together with values obtained by Voyager 1 radio occultation² (ingress, circles; egress, crosses). HASI pressure values are determined with an uncertainty of 1% along the entire descent.

potential temperature, the convective planetary boundary layer had a thickness of about 300 m at the place and time of landing.

Atmospheric electricity

Models of Titan's ionosphere predicted that galactic cosmic rays would produce an ionospheric layer with a maximum concentration of electrons between 70 and 90 km altitude^{34–37}. The Permittivity, Wave and Altimetry package²⁰ (PWA) measured the electrical state of the atmosphere below 140 km. We found that the electrical conductivity peaks at ~ 60 km. We might have seen evidence for lightning.

Observations of the electron and ion conductivities were made with two different techniques: relaxation and mutual impedance probes. The results of the relaxation probes (shown in Fig. 7a, b) indicate peaks in the electron/negative-ion conductivities at 60 km. Figure 7c shows that the altitude of the maximum in the conductivity (60 km) is confirmed by the mutual impedance probe measurements. This instrument gives the impedance of the medium at 45 Hz and yields a phase shift, which is sensitive to the presence of electrons only. The quadrupole probe also records the spectrum of the electric signal induced in the probe environment by the 45 Hz stimulus, in the bandwidth 0–9.22 kHz (Fig. 8a, active mode).

The electric field due to natural wave emissions was investigated during the descent, using the receiving dipole of the mutual impedance probe in two frequency ranges, 0–11.5 kHz and 0–100 Hz (Fig. 8b, c, passive mode). This provided a unique opportunity to investigate *in situ* lightning and related phenomena (for example, corona discharges) on Titan²¹ that would produce electromagnetic waves³⁸, excite global and local resonance phenomena in the surface–ionospheric cavity^{39,40} and could drive a global electric circuit²². Several impulsive events have been observed during the descent, for example at 2,800 s. The narrow-band wave emission seen near 36 Hz is reminiscent of a possible resonance generated by lightning activity in the spherical waveguide formed by the surface of Titan and the inner boundary of its ionosphere, but should be interpreted with caution. A comparison of the records presented in Fig. 8a and b shows that the first spectrogram (active mode) not only displays the

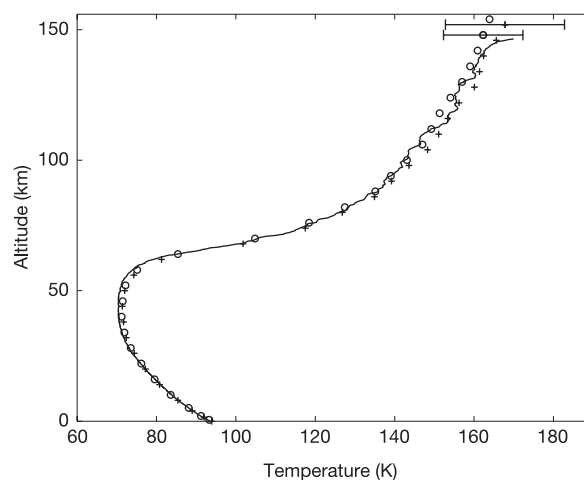


Figure 5 | Temperature profile of the lower atmosphere as measured by the temperature sensors, TEM²⁷ (expanded from Fig. 2). Temperature uncertainty is ± 0.25 K in the range from 60 to 110 K, and ± 1 K above 110 K. The temperature minimum of 70.43 K is reached at the tropopause (about 44 km; 115 ± 1 hPa). HASI temperatures are in very good agreement (within the error bars) with data obtained by Voyager radio occultation² (ingress, circles; egress, crosses) assuming a pure nitrogen atmosphere. The error bars for Voyager data are reported: ± 15 K (egress) ± 10 K (ingress) near the 200-km level, ± 0.5 K at the tropopause. At the tropopause, HASI measured temperature values ~ 1 K colder than Voyager², but reanalysis of these data³ suggested a similar temperature value (70.5 K) assuming a stratospheric composition of 98.5% N_2 plus 1.5% CH_4 .

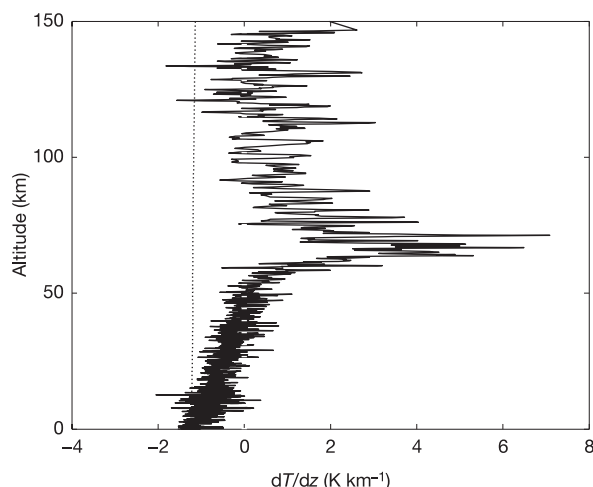


Figure 6 | The temperature lapse rate for the low atmosphere (expanded from Fig. 3). A number of inversion layers in the lower stratosphere and the strong temperature increase with altitude between 80 and 60 km are visible. Features present between 50 and 150 km could be related to turbulence due to Kelvin–Helmholtz instability induced by the large vertical shear of the wind speed, measured by the Doppler Wind Experiment³³. The temperature gradient in this part of the atmosphere has been derived from direct temperature measurements with vertical spatial resolution of the order of 200–150 m above 60-km altitude, and decreasing from 70 m down to 11 m until the last kilometre.

signals seen in the second spectrogram (passive mode), but also includes a broadband emission in the altitude range 110–80 km, and to lesser extent at altitudes lower than 25 km. It is believed that the energy injected in the medium at 45 Hz is partly dissipated in

nonlinear effects, which seems to strengthen the evidence for the presence of free charges in the upper atmosphere.

Surface

Before the probe landed, the nature of the surface was unknown. From the abundance of methane in its atmosphere, there was speculation that Titan might be covered by a methane ocean⁴¹, but recent observations¹⁴ have restricted the fraction of the surface covered with liquid to be just a few per cent. The probe touched down on a solid surface, which has properties something like wet sand⁴². The instruments continued to monitor the meteorological conditions for almost half an hour after impact.

The nature of Titan's surface at the landing site was investigated by spectral analysis of the Huygens radar return signal, the recording of the impact signature, *in situ* measurements of the ground electrical properties, and the surface environmental conditions.

The piezoresistive accelerometers of HASI recorded the impact instant at $T_0 + 2\text{h}27\text{min}49.840\text{ s}$ (where T_0 is the time of the parachute deployment device firing and corresponds to the beginning of the descent phase), when the event exceeded the threshold of $\sim 40\text{ m s}^{-2}$. A complete trace of the impact in the three orthogonal reference axes is shown in Fig. 9. The initial small peak in the X accelerometer data preceding the impact of the main probe may be related to a touch down on uneven topography, or the possible initial contact of a portion of the probe foredome, given the likely probe tilt at landing⁴². A sharp drop in acceleration is seen briefly in all three sensors at 8,869.86 s. The peak probe deceleration measured is 141 m s^{-2} , in reasonable agreement with the value measured by the accelerometer of the Science Surface Package (SSP)⁴². Over the length of the full data set, two possible events are seen in all three axes, at impact ($\sim 8,869.86\text{ s}$) and $\sim 3\text{ s}$ later at $\sim 8,872.2\text{ s}$. These correspond respectively to the initial impact event, and then to some short-term settling that may be surface related, or probe related (parachute system dynamics or structural relaxation of foredome). Further

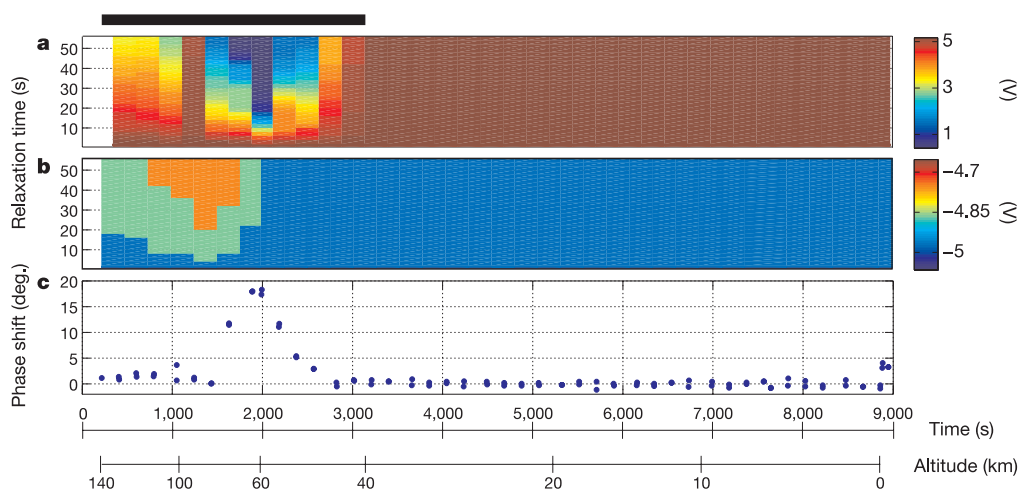


Figure 7 | A synopsis of PWA data: the signature of the ionosphere. The approximate extent of the ionized layer due to the interaction of cosmic rays with the atmosphere is indicated by a thick black line along the top axis. **a, b**, Relaxation carpets for $\Phi_0 = +5\text{ V}$ and -5 V , respectively. The relaxation probe, initially biased at a potential Φ_0 with respect to the vehicle body, subsequently returns to its equilibrium potential, Φ_∞ , with a time constant that yields the d.c. conductivity of the charges with polarity opposite to that of $\Phi_0 - \Phi_\infty$. The measurements taken during each relaxation cycle form a string of pixels aligned with the ordinate axis; voltages are given by the colour scales shown on the right-hand side. The electrode potential is measured every 20 ms during the first second, then every 2 s for the remainder of each 1 min cycle. These panels give a visual impression of the speed at which the potential of a conductive body (colour

coded) returns from $\pm 5\text{ V}$ to zero ('relaxes'), owing to the collection of ambient charges with opposite polarities. In the lower altitude range, for example, the colour of the carpet is uniform (brown for $+5\text{ V}$ and blue for -5 V), which shows that the ambient charge densities are low. Above 40 km, on the contrary, the distinctive carpet patterns tell us that the probe voltage is strongly affected by the ionized environment. **c**, Mutual impedance phase shift, $\Delta\phi = \phi_0 - \phi$ (non-calibrated). The a.c. conductivity is measured with a quadrupolar array. A current I with frequency 45 Hz and amplitude $\sim 10^{-10}\text{ A}$ is injected between two transmitting electrodes, and the voltage V induced between two receiving electrodes 2 m apart is measured. If the phase of V/I at 45 Hz is ϕ_0 in a vacuum and ϕ in a collisional medium, then the conductivity of the medium is proportional to $\tan(\phi_0 - \phi)$.

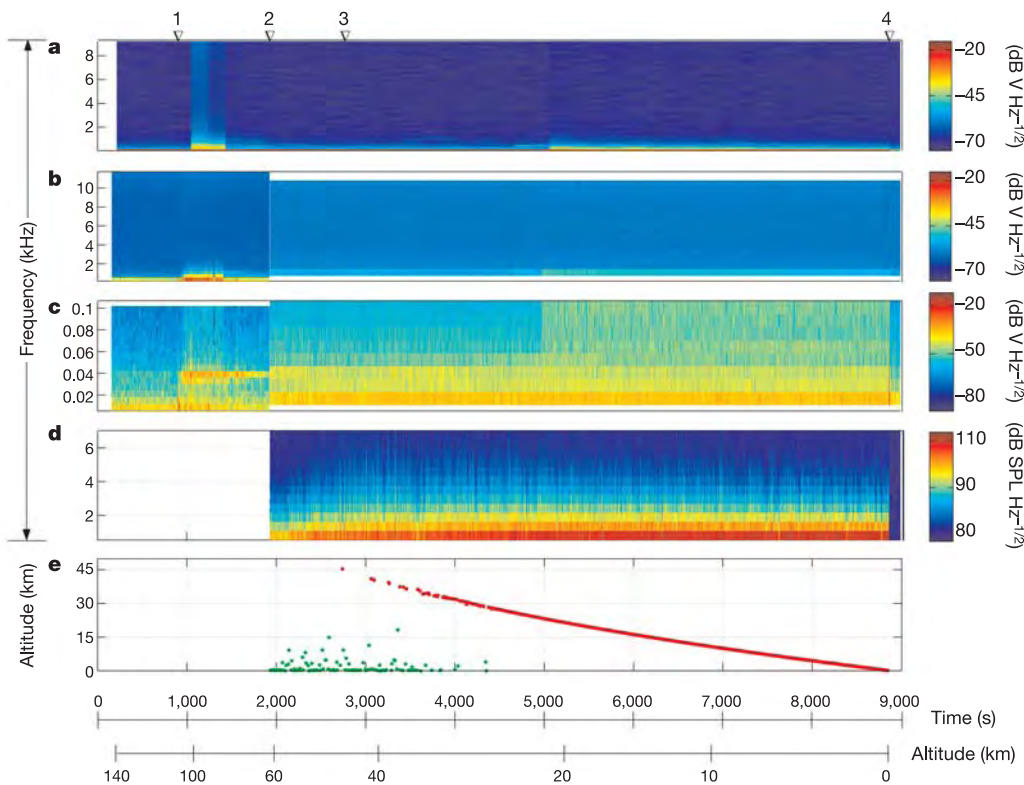


Figure 8 | A synopsis of PWA data: electric field, acoustic pressure and radar measurements. **a**, Dynamic spectrum of the voltage V measured between two electrodes 2 m apart, in the bandwidth 0–9.22 kHz, when a current stimulus I is injected between two transmitting electrodes. The spectrum of the signal provides information about its energy distribution as a function of frequency, at a given time. Successive spectra are represented by adjacent strings of pixels aligned with the ordinate axis, where spectral amplitude is coded in colour according to the logarithmic scale shown on the right-hand side. **b**, Dynamic spectrum of the voltage V measured with two electrodes 2 m apart, in the bandwidth 0–11.5 kHz, without current stimulus. **c**, Same as **b**, but in 0–100 Hz bandwidth. **d**, Dynamic spectrum of

acoustic differential pressure in the bandwidth 0–6.7 kHz. A sound pressure level (SPL) of 0 dB corresponds to $20 \mu\text{Pa}$. The variability of the acoustic noise is caused by changes in the atmospheric density and wind velocity⁵⁰. **e**, The altitude represented by the red dots is measured whenever the Radar Altimeter (RA) is locked on the surface; permanent lock is maintained from 34 km down to 150 m. At higher altitudes, the green dots indicate the distances at which the signal is returned by the atmosphere. Several events are identified with triangles along the top axis: (1) stabilizer parachute opening, (2) mode change, (3) impulsive event in **b**, (4) surface touch down. Discontinuities in time or frequency are artefacts due to mode change.

modelling of the probe structure behaviour is required to quantify these effects. Additionally, the area of stable data points immediately following the initial impact (8,870.1–8,870.3 s) may be due to a small bounce of the probe or to some structure vibrations. Integration of Y and Z axes after further processing, in combination with other sensors, will indicate any possible probe lateral movement. The integration of the accelerometer data gives a probe impact velocity of 4.33 m s^{-1} , in reasonable agreement with the values obtained by SSP⁴² and from the velocity profile during the last kilometre of the descent as derived from pressure measurements. For the final rest position of the probe, the X servo accelerometer gives an estimate of the probe tilt of about 11° , in good agreement with SSP tilt sensors.

At the surface, the HASI temperature and pressure sensors monitored the meteorological conditions for almost half an hour after the impact, measuring a temperature of $93.65 \pm 0.25 \text{ K}$ and a pressure of $1,467 \pm 1 \text{ hPa}$. The complex permittivity of the surface material is measured after impact with the PWA mutual impedance probe⁴³, at five frequencies. As a first estimation, the mean relative permittivity within the sensor range (radius 1 m, depth 2 m) is of the order of 2, in reasonable agreement with the measurements performed with the radar on board Cassini¹⁹.

In addition to providing altitude (Fig. 8), the Radar Altimeter measures the signal backscattered within the footprint of the beam, whose diameter is 0.14 times the altitude. This signal is strong and smooth with small variations over the ground track, indicating a

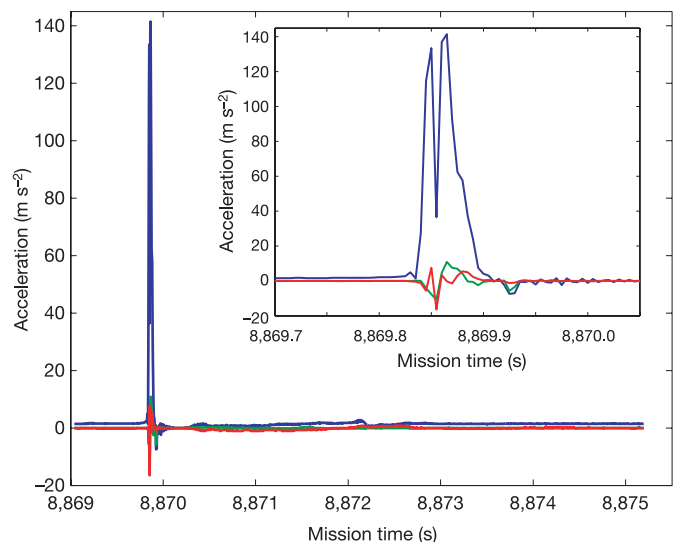


Figure 9 | The HASI signature of the impact trace, at 200 samples per second. The complete impact trace (6 s) is shown; the inset shows a magnified view of the deceleration peak. The X sensor (blue line) is aligned to the probe symmetry axis, corresponding to the descent direction. The Y (green line) and Z (red line) sensors are mounted orthogonal to the Huygens probe symmetry axis.

surface with little relief. The atmosphere was scanned and return signal from droplets was searched for, but no significant signature of rain could be found.

Discussion

Although the HASI data have now provided a great wealth of information on the conditions in the atmosphere and at the surface of Titan, many questions and challenges remain.

Atmospheric structure. The HASI temperature profile in the lower atmosphere was compared to the separate egress and ingress profiles based on the Voyager occultation experiment 25 years earlier. This comparison suggests that the atmosphere of Titan in mid-latitudes is uniform and slowly changing in accordance with model predictions. The open question is the poleward extent of this non-variability, given the latitudinal temperature gradient in the stratosphere inferred from infrared data^{44,45}. One interpretation of the south polar clouds is that they are due to heating associated with polar summer warming^{46,47}. If this is true, then the temperature profile in the polar summer should be different from the mid-latitude profiles sampled here, and could be revealed by Cassini infrared mapping⁴⁵ and radio science⁴⁸. In the middle and upper atmosphere (above 300 km), the prominent wave-like structure reported here requires further modelling to unambiguously identify the causal mechanisms. The observed vertical variation would suggest that large-scale temperature gradient in this region is also time variable. Unfortunately, the necessary observations of the time and spatial evolution of these structures must await future missions.

The atmosphere was scanned by the radar altimeter (before getting in lock), but no significant signature of rain was found. The instrument sensitivity to mass loadings of methane or other hydrocarbon droplets needs to be determined so that an upper limit to droplet mass loadings can be estimated.

Atmospheric electricity. The maximum in the conductivity due to positive ions, 20 km above the peak electron conductivity at 60 km, demands the presence of sufficient aerosols or electrophilic species in order to preserve charge neutrality. The altitude of the maximum conductivity due to electrons lies below that predicted by theoretical models^{35–37}. Several pulses similar to terrestrial sferics (natural electromagnetic waves) have been observed during the descent. Large convective clouds were observed near the south pole during the summer season¹⁷ and low-frequency electromagnetic waves could easily propagate from the south pole to the Huygens location. Lightning activity would also be consistent with the observations of waves in the Schumann frequency range.

Nature of the surface. The lack of any rhythmic motion during the half hour of operation on the surface indicated that the probe had landed on a solid surface rather than a liquid, which agrees with the image taken after the landing⁴⁹. The measured relative permittivity (of the order of 2) constrains the soil composition. No evidence for the presence of liquid phase on the surface was returned by the signal of the radar altimeter.

The HASI measurements of the atmospheric structure, electrical state and surface properties provide a unique insight into Titan's characteristics, unequalled in any planetary atmosphere except the Earth's. The many discoveries and puzzles will require synergetic analysis with the Cassini orbiter observations and years of laboratory and modelling efforts to solve.

Received 28 May; accepted 11 October 2005.

Published online 30 November 2005.

- McKay, C. P., Pollack, J. B. & Courtin, R. The thermal structure of Titan's atmosphere. *Icarus* **80**, 23–53 (1989).
- Lindal, G. F. *et al.* The atmosphere of Titan—an analysis of the Voyager 1 radio occultation measurements. *Icarus* **53**, 348–363 (1983).
- Lellouch, E. *et al.* Titan's atmosphere and hypothesized ocean: a reanalysis of the Voyager 1 radio-occultation and IRIS 7.7 μm data. *Icarus* **79**, 328–349 (1989).
- Vervack, R. J., Sandel, B. R. & Strobel, D. F. New perspectives on Titan's upper atmosphere from a reanalysis of the Voyager 1 UVS solar occultations. *Icarus* **170**, 91–112 (2004).
- Coustenis, A. *et al.* Titan's atmosphere from ISO mid-infrared spectroscopy. *Icarus* **161**, 383–403 (2003).
- Hubbard, W. B. *et al.* Results for Titan's atmosphere from its occultation of 28 Sagittarii. *Nature* **343**, 353–355 (1990).
- Sicardy, B. *et al.* The structure of Titan's stratosphere from the 28 Sgr occultation. *Icarus* **142**, 357–390 (1999).
- Tracadas, P. W., Hammel, H. B., Thomas-Osip, J. E. & Elliot, J. L. Probing Titan's atmosphere with the 1995 August stellar occultation. *Icarus* **153**, 285–294 (2001).
- Bouchez, A. H. *et al.* Adaptive optics imaging of a double stellar occultation by Titan. *Bull. Am. Astron. Soc.* **34**, 881 (2002).
- Sicardy, B. *et al.* The two stellar occultations of November 14, 2003: revealing Titan's stratosphere at sub-km resolution. *Bull. Am. Astron. Soc.* **36**, 1119 (2004).
- Lellouch, E., Hunten, D., Kockarts, G. & Coustenis, A. Titan's thermosphere profile. *Icarus* **83**, 308–324 (1990).
- Yelle, R. V. Non-LTE models of Titan's upper atmosphere. *Astrophys. J.* **383**, 380–400 (1991).
- Yelle, R. V., Strobel, D. F., Lellouch, E. & Gautier, D. *Engineering Models for Titan's Atmosphere* 243–256 (ESA SP-1177, European Space Agency, Noordwijk, 1997).
- Campbell, D. B., Black, G. J., Carter, L. M. & Ostro, S. J. Radar evidence for liquid surfaces on Titan. *Science* **302**, 431–434 (2003).
- Coustenis, A. *et al.* Maps of Titan's surface from 1 to 2.5 μm . *Icarus* **177**, 89–105 (2005).
- Meier, R., Smith, B. A., Owen, T. C. & Terrile, R. J. The surface of Titan from NICMOS observations with the Hubble Space Telescope. *Icarus* **145**, 462–473 (2000).
- Porco, C. C. *et al.* Imaging of Titan from the Cassini spacecraft. *Nature* **434**, 159–168 (2005).
- Brown, R. H. *et al.* Cassini Visual and Infrared Mapping Spectrometer (VIMS): Results for the SOI- and near-SOI period of the Cassini orbital tour. *Astron. Astrophys.* (submitted).
- Elachi, C. *et al.* Cassini radar views the surface of Titan. *Science* **308**, 970–974 (2005).
- Grard, R. *et al.* An experimental investigation of atmospheric electricity and lightning activity to be performed during the descent of the Huygens probe onto Titan. *J. Atmos. Terr. Phys.* **57**, 575–578 (1995).
- Desch, S. J., Borucki, W. J., Russell, C. T. & Bar-Nun, A. Progress in planetary lightning. *Rep. Prog. Phys.* **65**, 955–997 (2002).
- Tokano, T., Molina-Cuberos, G. J., Lammer, H. & Stumptner, W. Modelling of thunderclouds and lightning generation on Titan. *Planet. Space Sci.* **49**, 539–560 (2001).
- Lebreton, J.-P. & Matson, D. L. The Huygens probe: science, payload and mission overview. *Space Sci. Rev.* **104**, 59–100 (2002).
- Fulchignoni, M. *et al.* The characterization of Titan's atmospheric physical properties by the Huygens Atmospheric Structure Instrument (HASI). *Space Sci. Rev.* **104**, 395–431 (2002).
- Zarnecki, J. C. *et al.* *In-Flight Performances of the Servo Accelerometer and Implication for Results at Titan* 71–76 (ESA SP-544, European Space Agency, Noordwijk, 2004).
- Lebreton, J. P. *et al.* An overview of the descent and landing of the Huygens probe on Titan. *Nature* doi:10.1038/nature04347 (this issue).
- Ruffino, G. *et al.* The temperature sensor on the Huygens probe for the Cassini mission: Design, manufacture, calibration and tests of the laboratory prototype. *Planet. Space Sci.* **44–10**, 1149–1162 (1996).
- Harri, A.-M. *et al.* Scientific objectives and implementation of the Pressure Profile Instrument (PPI/HASI) for the Huygens spacecraft. *Planet. Space Sci.* **46**, 1383–1392 (1998).
- Mäkinen, T. Processing the HASI measurements. *Adv. Space Res.* **17**, 217–222 (1996).
- Niemann, H. B. *et al.* The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe. *Nature* doi:10.1038/nature04122 (this issue).
- Hinson, D. P. & Tyler, G. L. Internal gravity waves in Titan's atmosphere observed by Voyager radio occultation. *Icarus* **54**, 337–352 (1983).
- Friedson, A. J. Gravity waves in Titan's atmosphere. *Icarus* **109**, 40–57 (1994).
- Bird, M. K. *et al.* The vertical profile of winds on Titan. *Nature* doi:10.1038/nature04060 (this issue).
- Molina-Cuberos, G. J., López-Moreno, J. J., Rodrigo, R. & Lara, L. M. Chemistry of the galactic cosmic ray induced ionosphere of Titan. *J. Geophys. Res.* **104**, 21997–22024 (1999).
- Borucki, W. J. *et al.* Predictions of the electrical conductivity and charging of the aerosols in Titan's atmosphere. *Icarus* **72**, 604–622 (1987).
- Borucki, W. J., Whitten, R. C., Bakes, E. L. O., Barth, E. & Tripathi, S. Predictions of the electrical conductivity and charging of the aerosols in Titan's atmosphere. *Icarus* (in the press).
- Molina-Cuberos, G. J., López-Moreno, J. J., Rodrigo, R. & Schwingenschuh, K. Capability of the Cassini/Huygens PWA-HASI to measure electrical conductivity in Titan. *Adv. Space Res.* **28**, 1511–1516 (2001).

38. Schwingenschuh, K. *et al.* Propagation of electromagnetic waves in the lower ionosphere of Titan. *Adv. Space Res.* **28**, 1505–1510 (2001).
39. Nickolaenko, A. P., Besser, B. P. & Schwingenschuh, K. Model computations of Schumann resonance on Titan. *Planet. Space Sci.* **51**, 853–862 (2003).
40. Morente, J. A., Molina-Cuberos, G. J., Portí, J. A., Schwingenschuh, K. & Besser, B. P. A study of the propagation of electromagnetic waves in Titan's atmosphere with the TLM numerical method. *Icarus* **162**, 374–384 (2003).
41. Lunine, J. I., Stevenson, D. J. & Yung, Y. L. Ethane ocean on Titan. *Science* **222**, 1229–1230 (1983).
42. Zarnecki, J. C. *et al.* A soft solid surface on Titan as revealed by the Huygens Surface Science Package. *Nature* doi:10.1038/nature04211 (this issue).
43. Hamelin, M. *et al.* Surface and sub-surface electrical measurement of Titan with the PWA-HASI experiment on Huygens. *Adv. Space Res.* **26**, 1697–1704 (2000).
44. Coustenis, A. & Bézard, B. Titan's atmosphere from Voyager infrared observations. IV. Latitudinal variations of temperature and composition. *Icarus* **115**, 126–140 (1995).
45. Flasar, F. M. *et al.* Titan's atmospheric temperatures, winds, and composition. *Science* **308**, 975–978 (2005).
46. Brown, M. E., Bouchez, A. H. & Griffith, C. A. Direct detection of variable tropospheric clouds near Titan's south pole. *Nature* **420**, 7995–7997 (2002).
47. Tokano, T. Meteorological assessment of the surface temperatures on Titan: constraints on the surface type. *Icarus* **173**, 222–242 (2005).
48. Kliore, A. J. *et al.* Cassini Radio Science. *Space Sci. Rev.* **115**, 1–70 (2004).
49. Tomasko, M. G. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* doi:10.1038/nature04126 (this issue).
50. Ksanfomaliti, L. V. *et al.* Acoustic measurements of the wind velocity at the Venera 13 and Venera 14 landing sites. *Sov. Astron. Lett* **8**(4), 227–229 (1982).

Acknowledgements We thank the following people for their contributions to the realization of the HASI experiment: A. Buccheri, R. DeVidi, and M. Cosi of Galileo Avionica, A. Aboudan, S. Bastianello and M. Fabris of CISAS, M. Chabassière of LPCE, V. Brown, J.M. Jeronimo and L.M. Lara of IAA, R. Hofe of IWF, A. Smit, L. Smit and J. Van der Hooke from RSSD-ESTEC, H. Jolly from the UK, R. Pellinen, G. Leppelmeier, T. Siili, P. Salminen from FMI, and at the Aerodynamics Laboratory of Helsinki University of Technology T. Siikonen and B. Fagerström. HASI has been realised and operated by CISAS under a contract with the Italian Space Agency (ASI), with the participation of RSSD, FMI, IAA, IWF, LPCE and PSSRI sponsored by the respective agencies: ESA, TEKES, CSIC, BM:BWK, CNES and PPARC. We also acknowledge the long years of work by some hundreds of people in the development and design of the Huygens probe. The Huygens probe is part of the Cassini-Huygens mission, a joint endeavour of the National Aeronautics and Space Administration (NASA), the European Space Agency (ESA) and the Italian Space Agency (ASI).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.F. (francesca.ferri@unipd.it).

LETTERS

A soft solid surface on Titan as revealed by the Huygens Surface Science Package

John C. Zarnecki¹, Mark R. Leese¹, Brijen Hathi¹, Andrew J. Ball¹, Axel Hagermann¹, Martin C. Towner¹, Ralph D. Lorenz², J. Anthony M. McDonnell¹, Simon F. Green¹, Manish R. Patel¹, Timothy J. Ringrose¹, Philip D. Rosenberg¹, Karl R. Atkinson¹, Mark D. Paton¹, Marek Banaszkiewicz³, Benton C. Clark⁴, Francesca Ferri⁵, Marcello Fulchignoni⁶, Nadeem A. L. Ghafoor⁷, Günter Kargl⁸, Håkan Svedhem⁹, John Delderfield¹⁰, Manuel Grande¹⁰, David J. Parker¹⁰, Peter G. Challenor¹¹ & John E. Geake^{12‡}

The surface of Saturn's largest satellite—Titan—is largely obscured by an optically thick atmospheric haze, and so its nature has been the subject of considerable speculation and discussion¹. The Huygens probe entered Titan's atmosphere on 14 January 2005 and descended to the surface using a parachute system². Here we report measurements made just above and on the surface of Titan by the Huygens Surface Science Package^{3,4}. Acoustic sounding over the last 90 m above the surface reveals a relatively smooth, but not completely flat, surface surrounding the landing site. Penetrometry and accelerometry measurements during the probe impact event reveal that the surface was neither hard (like solid ice) nor very compressible (like a blanket of fluffy aerosol); rather, the Huygens probe landed on a relatively soft solid surface whose properties are analogous to wet clay, lightly packed snow and wet or dry sand. The probe settled gradually by a few millimetres after landing.

The Surface Science Package (SSP) comprises nine independent sensors. An in-depth technical description has been given in earlier papers^{3,4}. The nine sensors were chosen such that some were designed primarily for landing on a solid surface and others for a liquid landing, with eight also operating during the descent. All sensors appear to have performed normally during the probe mission. Those sensors intended for a liquid landing scenario (refractometer, permittivity and density sensors) would have performed correctly for a liquid landing case; they are still under analysis for any secondary results. The SSP science data were redundantly transmitted on the two communication chains so that the loss of data on chain A did not result in any data loss for the SSP².

The Acoustic Properties Instrument–Sonar (API-S) recorded the approach to the surface on final descent (Fig. 1). API-S is a pulse send–receive sonar, where the time of flight gives distance (and hence final descent speed). The probe vertical speed just before landing was determined as $4.60 \pm 0.05 \text{ m s}^{-1}$. The peak width and signal strength are influenced by surface topography, probe position and acoustic reflectivity according to the usual radar equation for an extended target.

As Huygens descended towards the surface the sensor footprint shrank, and a smaller area of terrain was illuminated. Owing to variation in probe tilt and wind drift during descent, the sensor illuminated different areas of ground for each pulse, with partial

overlap. Initial derivation of surface acoustic reflectivity shows no significant variation as a function of altitude, implying that the landing site as seen by Huygens is typical of the local surroundings (the maximum area sampled by API-S, for the highest altitude of around 90 m, is approximately a circle of 40 m diameter).

For all returns the peak widths are typically 30–50 ms wide, showing no trends. This implies that the surface is topographically

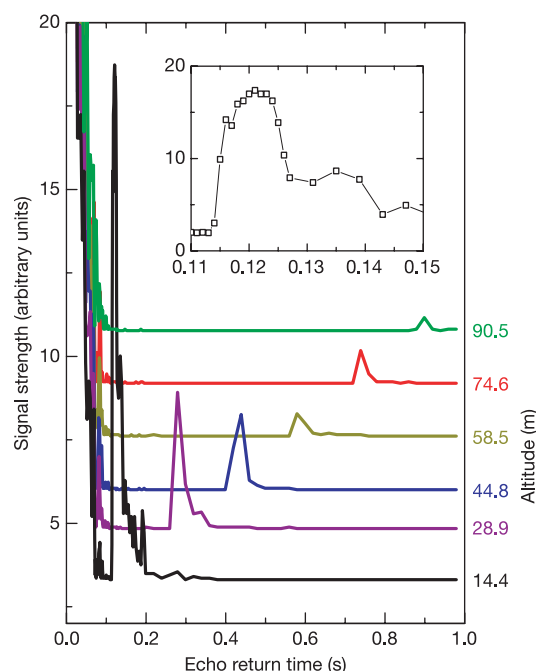


Figure 1 | Acoustic sonar (API-S) surface echoes. Note that the larger signals to the left of the plot are the result of the sensor ringing from the send pulse intruding into the receive time window. The inset is a zoom on the final API-S surface detection from 14.4 m altitude (at the time of pulse transmission). A speed of sound measurement of $191.9 \pm 1.8 \text{ m s}^{-1}$ from the SSP Acoustic Velocity (API-V) sensors near the surface is used to convert ranging time delay into altitude.

¹The Open University, Walton Hall, Milton Keynes MK7 6AA, UK. ²University of Arizona, Lunar and Planetary Laboratory, Tucson, Arizona 85721, USA. ³Polish Academy of Sciences, Ul Bartycka 18 A, Warszawa, PL-00716, Poland. ⁴Lockheed Martin Astronautics, PO Box 179, Denver, Colorado 80201, USA. ⁵CISAS G. Colombo, University of Padova, Via Venezia 15, 35131 Padova, Italy. ⁶LESIA, Paris Observatory, 5 Place Janssen, 92195, Meudon, France. ⁷MD Robotics, 9445 Airport Road, Brampton, Ontario, L6S 4J3, Canada. ⁸Space Research Institute, Austrian Academy of Sciences, Schmiedlstraße 6, A-8042 Graz-Messendorf, Austria. ⁹ESA/ESTEC, Research and Scientific Support Department, Postbus 299, 2200 AG, Noordwijk, The Netherlands. ¹⁰Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, UK. ¹¹Southampton Oceanography Centre, Empress Dock, Southampton SO14 3ZH, UK. ¹²Physics Department, UMIST, Manchester M60 1QD, UK. [‡]Deceased.

similar over all sampled beam footprints. However, this width is greater than would be expected for a purely flat surface, implying that some small-scale vertical topography is present.

The final peak immediately before impact is at a height of 14.4 m at the time of pulse transmission, with a beam footprint of $\sim 26 \text{ m}^2$ (equivalent to a circle of $\sim 2.9 \text{ m}$ radius). This final peak is recorded by the SSP at higher time resolution than previous ones, giving more information on surface structure (Fig. 1 inset). The relatively broad shape of this peak indicates that the surface cannot be completely flat, or concave over the footprint. However, the flat top of the peak also requires that there be some local height variation over the surface sampled within the footprint. Rock size determined from the post-landing surface images will provide a good starting point for further collaborative analysis. When averaged to a lower time resolution, the width of the final peak is entirely comparable to the width of the higher-altitude peaks, implying that they are seeing very similar terrain.

Together these data suggest a surface that is relatively flat but not completely smooth; such an interpretation is compatible with the Descent Imager and Spectral Radiometer (DISR) surface images, suggesting that perhaps the DISR images show a typical surface that probably surrounds the probe in all directions. The fact that slight horizontal and vertical topographic variation is seen over the footprints, rather than a completely flat plain, implies a certain level of complexity during the history of surface formation in the region of the landing site.

There is a well-established history of determining the mechanical properties of a planetary surface from the dynamics of a spacecraft landing⁵. Titan's outer surface layers were expected to be dominated by water ice and organic materials, although other ices and minerals could not be ruled out. As with other planetary surfaces, these materials might be processed by impacts, volcanism and erosion¹. The SSP includes an impact penetrometer and an accelerometer to measure mechanical properties of the surface material at the landing site. The ACC-I accelerometer is a single-axis piezoelectric accelerometer able to produce a successful measurement for all survivable landing scenarios, but for the hardest surfaces it was more likely that the structures of the probe would have been crushed, damping the measured deceleration. For such cases the impact penetrometer (ACC-E) was provided to measure directly the penetration resistance of the ground^{5,6}, yielding strength and texture information through a piezoelectric force transducer positioned behind a 16-mm-diameter hemispherical tip.

ACC-I and ACC-E together covered the wide range of properties that could have been encountered, from liquids or very soft material to solid, hard ice. Their ranges of applicability overlapped—for intermediate-strength materials the surface would be soft enough not to crush the probe, and thus produce meaningful output from the accelerometer, yet hard enough also to produce a meaningful signal from the penetrometer. Also measuring the impact dynamics were SSP's two-axis tilt sensor (TIL) and the three piezoresistive (PZR) accelerometers of the Huygens Atmospheric Structure Instrument Accelerometer (HASI ACC) experiment subsystem⁷.

Figure 2 shows the impact signatures from the penetrometer and accelerometers. The impact triggered ACC-E at a mission time² of $T_0 + 8,869.7598 \text{ s}$ as the penetrometer tip penetrated the surface, followed by ACC-I, which triggered at $T_0 + 8,869.7695 \text{ s}$ as the probe foredome struck the surface (T_0 is defined as the start of the descent sequence at 09:10:21 UTC). This event also seems to have caused the broad peak in the ACC-E signal following the period of 'clean' penetration. The observations that ACC-E triggered with a signal containing detailed structure and ACC-I returned a signal of brief duration with minimal rebound immediately ruled out a liquid landing.

The raw signal from ACC-E has been processed to correct for the transfer function of the electronics and digitization noise at the analog-to-digital converter (ADC). The processed, calibrated signal

(Fig. 3) shows the following features: a shallow rise at the start of the event; a strong peak; a smooth plateau at around 50 N, and a broad, smooth peak once the measurement is disrupted by impact of the probe's foredome. The near-constant force of 50 N over its $\sim 2 \text{ cm}^2$ projected area gives a dynamic penetration resistance of 250 kPa. Terrestrial materials with these strength characteristics include lightly packed snow, tar, and wet sand or clay. Initial results from laboratory experiments using an identical penetrometer striking at 3.7 m s^{-1} (the maximum velocity currently achievable with our test rig) into a range of room-temperature analogue targets suggest that the signal is consistent with an ACC-E impact into a moderately firm, perhaps wet granular material overlain by an ice pebble or—perhaps less likely, given the prevalence of pebbles and cobbles in the DISR surface images⁸—a thin crust, and in either case coated with a very soft top layer. The signal's subsurface plateau phase shows a lack of prominent positive-going short-period structure, yet is not completely smooth, indicating the presence of some small-scale texture. This is consistent with the penetrometer encountering a mixture that is likely to be poorly sorted but containing nothing coarser than sand, granules and small pebbles (as defined by the Udden–Wentworth scale⁹). Some inhomogeneity, including voids, may also be present. The slight downward trend in the plateau phase could be consistent with the presence of liquid among the grains and a liquid content increasing with depth.

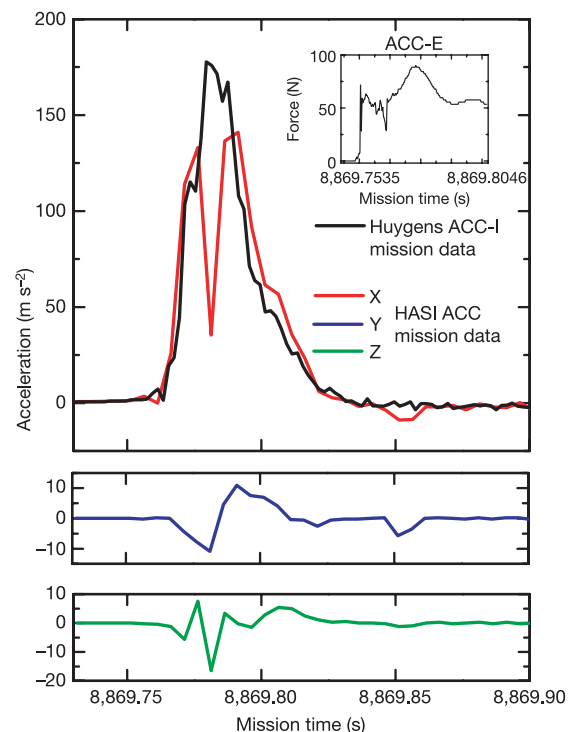


Figure 2 | Impact deceleration profiles. Main panels, SSP ACC-I (black) and HASI ACC PZR X, Y, Z (red, blue, green) accelerometer impact signatures, with SSP ACC-E penetrometer impact signature (inset). ACC-I (a single-axis piezoelectric accelerometer) and PZR X (a piezoresistive accelerometer) are both aligned parallel to the probe's axis, while PZR Y, Z are perpendicular, to measure transverse accelerations. Note that the PZR data has been time-shifted to match the ACC-I impact time; this is within the 125 ms uncertainty between the two experiments. The ACC-I accelerometer (an Endevco 2271 AM20) is aligned with the probe's axis of symmetry (X axis) but mounted 0.325 m from it, on the SSP electronics box. It was sampled at 500 Hz with 12 bit resolution over the range $\pm 90g$ for a duration of 512 samples. The HASI PZR sensors (Endevco type 7264A-2000T) were mounted close to the probe's centre of mass. They had an absolute accuracy of $\pm 4 \text{ m s}^{-2}$ and resolution of 0.15 m s^{-2} , and were sampled at 200 Hz.

Both ACC-I and the (parallel) HASI ACC PZR X sensor (see Fig. 2 legend for definitions of PZR X, Y and Z) registered a small precursor peak of a few m s^{-2} in amplitude. Although it is tempting to associate this with the impact of the ACC-E penetrometer, the peak is an order of magnitude larger than would be obtained from the peak force actually measured by ACC-E. One possibility is that it may be due to impact of the probe's foredome with an isolated protuberance such as an 'ice cobble' resting on the surface.

The peak decelerations parallel to the probe's axis seen by ACC-I and PZR X were 178 m s^{-2} and 141 m s^{-2} , respectively. The cause of the double peak structure of the PZR X signal is currently under investigation; it might be due to a high-frequency resonance of the experiment platform being sampled by the HASI electronics at a lower frequency of 200 Hz, rather than real probe dynamics. ACC-I showed little if any such dip—if platform resonance were responsible for the dip, then this could be explained by the location of ACC-I closer to the vibrational node around the platform's edge. After initial transfer function processing and integration, the ACC-I signal gives a speed change of 4.63 m s^{-1} with minimal bounce. This is corroborated by the PZR X data, which give a peak of 4.33 m s^{-1} . Further integration of the accelerometry leads to an estimate of the distance over which the probe decelerated of 0.12 m. The radius of curvature of the probe foredome is 1.215 m, thus the maximum contact area (assuming a 0.12 m penetration) is 0.92 m^2 . The deceleration peaks at a distance of 0.09 m, at which point the maximum contact area would be approximately 0.69 m^2 . A deceleration of the 200.5 kg probe of 178 m s^{-2} requires a force of 36 kN. Exerted over the contact area indicated above, this implies a dynamic penetration resistance of $>52 \text{ kPa}$. This is a tighter constraint than that implied by the

persistence of $\sim 0.1\text{-m}$ -scale cobbles (in DISR surface images⁸), supported by the bulk surface material which implies bearing strength of $>\sim 0.1 \text{ kPa}$. The magnitudes of these accelerations also rule out the presence of a very hard consolidated material down to the depth penetrated.

The difference between the penetrometer and accelerometer determinations may indicate some structural filtering by the probe, or that the penetrometer may have struck a site harder than the average beneath the probe. Such a discrepancy may not be surprising, given the presence of the $\sim 0.1\text{-m}$ -scale cobbles. The accelerometer and penetrometer results may be reconciled if the probe, instead of landing directly on the 250-kPa-strength substrate sampled by the penetrometer, crashed onto ice cobbles with a collective area of $\sim 0.2 \text{ m}^2$. These cobbles in turn pushed into the substrate, acting as 'penetrometers' themselves with a smaller area than the foredome, and thus yielding the modest observed deceleration.

Preliminary evaluation by comparison with scale models¹⁰ and numerical simulations¹¹ suggests that the surface was neither hard (like solid ice) nor very compressible (like a blanket of fluffy aerosol). Analogue materials with mechanical properties consistent with the data include tarry materials like wet clay, somewhat cohesive material like lightly packed snow, and wet or dry sand. Interpretation is complicated by the possible presence of cobbles resting on the surface (DISR), which would change the effective shape of the penetrating probe (from the case of a flat surface).

The probe's apparent tilt (that is, the angle between the probe's axis and the tiltmeter's instantaneous acceleration vector) just before landing was around 9° , but had been varying in the range $4\text{--}16^\circ$ during the 20 s before impact, on timescales at least as short as the 1 s sampling interval. By averaging the signal over the final 1 km of descent to smooth out the dynamic effects of probe motion, it is clear that there existed a mean tilt of about 8° . Oscillations are visible in the TIL signal until some 10 s after landing. Damped oscillations within the TIL sensors may offer only a partial explanation for this, suggesting that the probe took several seconds to come to a final rest. The probe's tilt after the impact event was very similar to that before impact, at 10.3° . This suggests that the surface material was deformed largely plastically (whether by shear or brittle failure and compression) and was readily penetrated.

TIL indicates a gradual change in the angle of the probe at a rate of 0.2° h^{-1} during the 70 min after landing for which data were received. This is corroborated by data from the HASI ACC X servo-accelerometer (the high resolution part of the HASI ACC experiment subsystem, co-aligned with the PZR X sensor), which shows a similar rate of change. The HASI data would give the same absolute magnitude of tilt as that from TIL if one assumed the value of gravity at a radius of $2,578.5 \pm 1 \text{ km}$, as compared with the published value² of $2,575 \pm 2 \text{ km}$, which gives an apparent tilt difference of $<0.5^\circ$. This is tentative, however, as measurements are taken at the limit of resolution of the sensors. This degree of settling amounts to a shift of a few millimetres in the probe's position.

Modelling¹, together with optical, radar and infrared spectrometer images from Cassini^{12–14} and images from the Huygens probe⁸ indicate a variety of possible processes modifying Titan's surface. These include tectonism, cryovolcanism, impacts and fluvial erosion. Fluvial and marine/lacustrine processes appear most prominent at the Huygens landing site, although aeolian activity cannot be excluded. Thus the SSP and HASI accelerometer impact dynamics data are consistent with two plausible interpretations for the soft substrate material: solid, granular material having either low or zero cohesion, or a fluid component. The mixture resulting from the latter possibility would be analogous to a wet sand or a textured tar/wet clay. These possibilities, between which our data alone cannot discriminate, would involve 'sand' made presumably of ice grains from impact or fluvial erosion, wetted by liquid methane, or a collection of photochemical products and/or fine-grained ice making a plastic or viscoplastic material, that is, a 'tar'.

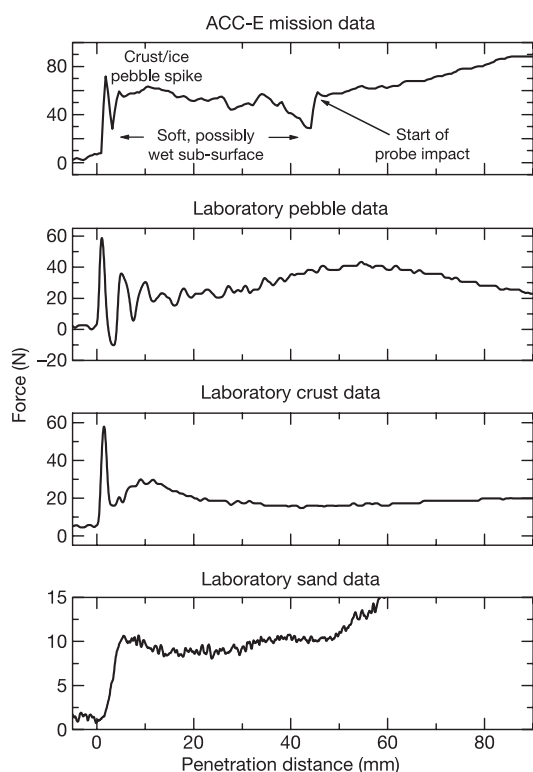


Figure 3 | A comparison of penetrometer force profiles for Titan and laboratory analogues. Top to bottom: SSP ACC-E mission data, and laboratory data for impact onto a pebble, impact onto a surface crust layer, and impact onto sand. ACC-E was mounted on a pylon that protruded below the probe foredome to give 55 mm of undisturbed penetration before the main structure of the probe contacted the surface. The force measurement employed pseudo-logarithmic amplification and was sampled at 10 kHz and 8 bit resolution with a range of approximately 6 kN (although structural failure would occur at around 2 kN) for a duration of 512 samples.

Received 2 June; accepted 6 September 2005.

Published online 30 November 2005.

1. Lorenz, R. D. & Lunine, J. I. Titan's surface before Cassini. *Planet. Space Sci.* **53**, 557–576 (2005).
2. Lebreton, J.-P. *et al.* An overview of the descent and landing of the Huygens probe on Titan. *Nature* doi:10.1038/nature04347 (this issue).
3. Zarnecki, J. C. *et al.* in *Huygens: Science, Payload and Mission* (ed. Wilson, A.) 177–195 (SP-1177, ESA Publications Division, Noordwijk, The Netherlands, 1997).
4. Zarnecki, J. C., Leese, M. R., Garry, J. R. C., Ghafoor, N. & Hathi, B. Huygens' Surface Science Package. *Space Sci. Rev.* **104**, 593–611 (2002).
5. Kömle, N. I., Kargl, G., Ball, A. J. & Lorenz, R. D. (eds) *Penetrometry in the Solar System* (Austrian Academy of Sciences Press, Vienna, 2001).
6. Lorenz, R. D. *et al.* An impact penetrometer for a landing spacecraft. *Meas. Sci. Technol.* **5**, 1033–1041 (1994).
7. Fulchignoni, M. *et al.* *In situ* measurements of the physical characteristics of Titan's atmosphere. *Nature* doi:10.1038/nature04314 (this issue).
8. Tomasko, M. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* doi:10.1038/nature04126 (this issue).
9. Wentworth, C. K. A scale of grade and classification terms for clastic sediments. *J. Geol.* **30**, 377–392 (1922).
10. Seiff, A. *et al.* Determination of physical properties of a planetary surface by measuring the deceleration of a probe upon impact: Application to Titan. *Planet. Space Sci.* **53**, 594–600 (2005).
11. Lorenz, R. D. Huygens probe impact dynamics. *ESA J.* **18**, 93–117 (1994).
12. Porco, C. C. *et al.* Imaging of Titan from the Cassini spacecraft. *Nature* **434**, 159–168 (2005).
13. Elachi, C. *et al.* Cassini radar views the surface of Titan. *Science* **308**, 970–974 (2005).
14. Sotin, C. *et al.* Release of volatiles from a possible cryovolcano from near-infrared imaging of Titan. *Nature* **435**, 786–789 (2005).

Acknowledgements We acknowledge the work of the SSP Team and the HASI Accelerometer Team both past and present in the design, build, test, calibration and operation of these experiments. This work has been funded by the UK Particle Physics and Astronomy Research Council, The Royal Society, the ESA, NASA, CNES and the Polish State Committee for Scientific Research.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.C.Z. (J.C.Zarnecki@open.ac.uk).

LETTERS

Complex organic matter in Titan's atmospheric aerosols from *in situ* pyrolysis and analysis

G. Israël¹, C. Szopa¹, F. Raulin², M. Cabane¹, H. B. Niemann³, S. K. Atreya⁴, S. J. Bauer⁵, J.-F. Brun¹, E. Chassefière¹, P. Coll², E. Condé⁶, D. Coscia², A. Hauchecorne¹, P. Millian⁷, M.-J. Nguyen², T. Owen⁸, W. Riedler⁹, R. E. Samuelson¹⁰, J.-M. Siguier⁷, M. Steller¹¹, R. Sternberg² & C. Vidal-Madjar¹²

Aerosols in Titan's atmosphere play an important role in determining its thermal structure^{1–3}. They also serve as sinks for organic vapours⁴ and can act as condensation nuclei for the formation of clouds^{5,6}, where the condensation efficiency will depend on the chemical composition of the aerosols^{5,7}. So far, however, no direct information has been available on the chemical composition of these particles. Here we report an *in situ* chemical analysis of Titan's aerosols by pyrolysis at 600 °C. Ammonia (NH₃) and hydrogen cyanide (HCN) have been identified as the main pyrolysis products. This clearly shows that the aerosol particles include a solid organic refractory core. NH₃ and HCN are gaseous chemical fingerprints of the complex organics that constitute this core, and their presence demonstrates that carbon and nitrogen are in the aerosols.

Although the Gas Chromatograph and Mass Spectrometer (GCMS) was primarily devoted to the analysis of atmospheric gases⁸, it was also used for the analysis of vaporized particulates through coupling to the Aerosol Collector and Pyrolyser (ACP) experiment. The ACP instrument^{9,10} collected two distinct atmospheric samples over separate altitude ranges (130–35 km and 25–20 km, respectively) during the Huygens probe descent (see Supplementary Information). The composition of each sample was analysed by the GCMS in three stages (see Table 1). First, the most volatile part of a given sample was analysed by the GCMS at 'ambient' collection temperature. Second, the remaining part of the sample was heated in the ACP oven to 250 °C in order to vaporize all volatile components of the collected aerosols, and analysis of the gaseous products was then carried out. Last, the remainder of the sample was subjected in the oven to a temperature of 600 °C. The high temperature in this last stage ensured that the refractory material composing the aerosol particles was thermally decomposed (pyrolysed) into molecular gaseous products. These products contribute to the composition of the gaseous sample to be analysed by GCMS. An in-depth technical description of the ACP experiment and its coupling with the GCMS has been given in earlier articles^{9–11}. Relevant temperatures and altitudes during descent are also given in Table 1.

In this report we focus on the composition of the refractory material making up the collected aerosols, and hence restrict our attention to the third stage of the analysis, in which the sample is heated to 600 °C before transferring the gaseous products of pyrolysis to the GCMS. Only data obtained using the direct Mass Spectrometry (MS) mode are analysed here. As explained in the companion paper⁸,

all data related to the gas chromatograph part of the GCMS will be reported later. In the obtained spectra, values of mass to charge ratio (m/z) above 50 are close to the noise level and are difficult to determine. We therefore restrict our analysis to the range $m/z = 2–50$.

Results from the transfer of the evolved gases to the GCMS experiment after heating the samples at 600 °C are compared with results for the reference background in Figs 1 and 2. Two of the most noticeable enhancements of signal for the transfer samples (that are not associated with molecular nitrogen) are those at $m/z = 17$ and 27.

Mass spectra measured during the transfer of the gases that evolved from the first aerosol sample (after pyrolysis at 600 °C) reveal a pyrolysis fragment feature at $m/z = 17$ (Fig. 1). According to the National Institute of Standards and Technology (NIST) library on mass spectrometry, this specific feature can be attributed to CH₃D, ¹³CH₄ or NH₃. A comparison of relative strengths of the features at $m/z = 16$ and $m/z = 17$ indicates that CH₃D must have no noticeable contribution at $m/z = 17$, given the known D/H ratio in Titan's atmosphere (about $(2.3 \pm 0.5) \times 10^{-4}$; ref. 12). ¹³CH₄ must contribute, but if the $m/z = 17$ signal was entirely due to this compound, our data would provide a ¹²C/¹³C ratio of 48. This is much lower than the value of 82.3 measured in the gas phase by the GCMS. It would then correspond to an unexplainable isotopic enrichment in ¹³C in the aerosols when compared with the gaseous phase, whereas it is expected that kinetically driven chemical processes favour ¹²C insertion, as already observed in laboratory experiments¹³. Consequently, the $m/z = 17$ signature must be attributed to another species, in combination with the contribution of ¹³CH₄. The only possible candidate for this contribution is NH₃. A similar interpretation can be made for the $m/z = 17$ feature in Fig. 2. In this case, the ¹²C/¹³C ratio is 56. Finally, further evidence can be seen in Fig. 3a and b, where the evolution of the signal at $m/z = 16$ and $m/z = 17$ respectively, during the transfer of the gases evolved from the second sample pyrolysed at 600 °C, strongly favours identification of ammonia.

When the second aerosol sample is pyrolysed at 600 °C (Fig. 2), the signal intensity increases, compared with the signal presented in Fig. 1, possibly owing to the collection of a larger number of particles; in addition, a pyrolysis fragment feature that was lost in the background in Fig. 1 appears in Fig. 2 at $m/z = 27$. The NIST library molecular fragment list suggests that several compounds may contribute to this feature at $m/z = 27$. These compounds are C₂H₄,

¹Service d'Aéronomie UMR 7620 CNRS, ISPL, Universities Paris 6 and Versailles-Saint Quentin, Verrières-le-Buisson, F 91371, France. ²Laboratoire Interuniversitaire des Systèmes Atmosphériques, UMR 7583 CNRS, Universities Paris 12 and Paris 7, Créteil, F 94010, France. ³NASA Goddard Space Flight Centre, Greenbelt, Maryland 20771, USA.

⁴University of Michigan, Ann Arbor, Michigan 48109, USA. ⁵Institute for Meteorology and Geophysics, University of Graz, Graz A 8010, Austria. ⁶Centre National d'Etudes Spatiales (CNES), Toulouse, F 31401, France. ⁷Office National d'Etudes et de Recherches Aéropatiales, Toulouse, F 31055, France. ⁸Institute for Astronomy, University of Hawaii at Manoa, Honolulu, Hawaii 96822, USA. ⁹Space Research Institute, Graz A 8010, Austria. ¹⁰Department of Astronomy, University of Maryland, College Park, Maryland 20742, USA. ¹¹Institut für Weltraumforschung, Österreichische Akademie der Wissenschaften, Graz A 8042, Austria. ¹²Laboratoire de Recherche sur les Polymères, UMR 7581 CNRS, Thiais, F 94320, France.

Table 1 | Results for the two aerosol samples collected and analysed by ACP

Parameters	Sample 1			Sample 2		
Sampling altitudes (km)	130–35			25–20		
Atmospheric pressure (mbar)	3–176			320–430		
Analysis altitudes (km)	32	30	27	16	15	13
Atmospheric pressure (mbar)	250	275	325	640	685	760
Sample temperature in the ACP oven (°C)*	'Ambient': –120 to –90	250	600	–74	250	600

Altitudes corresponding to sampling and analysis were provided by the HASI experiment. All samples were analysed using direct mass spectrometry measurements with the Huygens GCMS instrument.

* Values are $\pm 10^\circ\text{C}$.

C_2H_6 , C_3H_8 and HCN. The main fragments of C_2H_4 , C_2H_6 and C_3H_8 detected by MS are observed at $m/z = 28$ and 29 according to the NIST library. As the atmospheric nitrogen, and the labelled molecular nitrogen $^{15}\text{N}^{15}\text{N}$ used in ACP, are strong contributors to the signal obtained at these m/z values, it is not possible to point out the presence of C_2H_4 , C_2H_6 and C_3H_8 with these features. However, if these compounds were contributing to the feature at $m/z = 27$, we should see a noticeable contribution of the C_2 hydrocarbons at $m/z = 26$ and noticeable contributions of C_3H_8 at $m/z = 39$, 41 and 43 . In the present case, the pyrolysis feature at $m/z = 26$ is much smaller than that at $m/z = 27$, and no features are evident at $m/z = 39$, 41 and 43 . We conclude (although it is not easy to verify this owing to the poor resolution of the Mass Spectrometer) that HCN is the main contributor to the feature at $m/z = 27$ in the same way that NH_3 is the dominant contributor to the feature at $m/z = 17$.

If these organic molecules had simply been condensed onto aerosol particles, they would have been driven off during the second analysis stage (at 250°C) and further cleaning of the oven. We can therefore conclude that the NH_3 and HCN observed in the third stage are pyrolysis products from the refractory aerosol material itself. This is confirmed by the fact that the GCMS experiment did not observe

NH_3 and HCN in its atmospheric sampling⁸. This ACP result, predicted by theoretical models⁴, is of prime importance, as it is the first evidence of the presence of complex macromolecular organic matter in Titan's atmosphere.

Possible chemical pathways proposed for the production of aerosols from the gaseous chemistry^{4,12} are: (1) polymerization of C_2H_2 ; (2) polymerization of nitriles; (3) formation of polyaromatic molecules; and (4) copolymerization (aliphatic and aromatic). Our data clearly indicate that nitrogen is incorporated into the chemical structure of Titan's aerosols, ruling out the possibility that these aerosols consist solely of polyacetylenes or other pure hydrocarbon compounds. To go further in the determination of Titan aerosol production pathways, one needs to rely on data obtained from experimental simulations—either by photochemistry^{14,15} or cold plasma discharges^{16–19}—that provide laboratory analogues (called

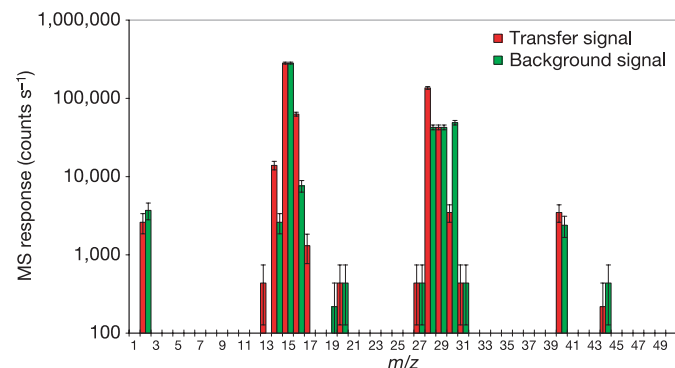


Figure 1 | Signal resulting from the mass spectrometry (MS) analysis of the gases evolved from pyrolysis of the first aerosol sample. Red, ion count rates per second versus mass per unit charge (m/z) measured during the transfer of the gases evolved from the first aerosol sample pyrolysed at 600°C , and green, its corresponding background. Mass scans from $m/z = 2$ to 141 are acquired in unit mass steps every 0.937 s. One count on the detector during the acquisition is equal to 217.7 counts s^{-1} . Labelled $^{15}\text{N}^{15}\text{N}$ is used to achieve the gas transfers from ACP to GCMS. Error bars represent $\pm$ one standard deviation. After completion of the heating cycle, and transfer of gaseous pyrolysis products to the GCMS for analysis, the ACP oven is vented to the atmosphere and the background gas measurements are undertaken by GCMS. Thus, atmospheric gases contribute to the background signal along with the main $^{15}\text{N}^{15}\text{N}$ carrier gas. The signals at $m/z = 14$, 15 and 16 can be mainly ascribed to methane, and to $^{14}\text{N}^{14}\text{N}$ (for $m/z = 14$) and $^{15}\text{N}^{15}\text{N}$ (for $m/z = 15$). The intensity of the feature at $m/z = 17$ increases from 0 counts s^{-1} (background) to $1,306$ counts s^{-1} (analysis), demonstrating that the molecule responsible is a pyrolysis product. Analysis of the $m/z = 17$ signal demonstrates a contribution of NH_3 in addition to that of $^{13}\text{CH}_4$.

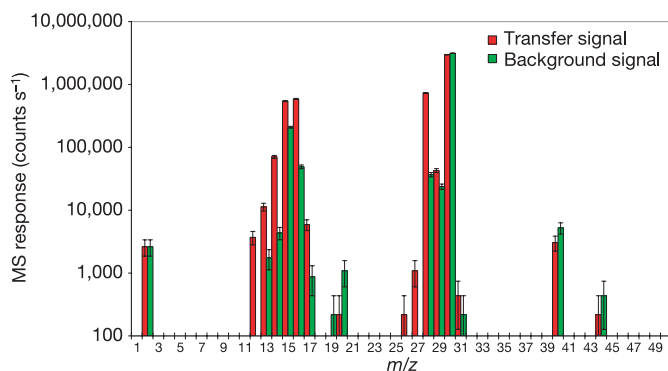


Figure 2 | Signal resulting from the MS analysis of the gases evolved from pyrolysis of the second aerosol sample. Red and green, as for Fig. 1 but for the second aerosol sample; error bars represent $\pm$ one standard deviation. In these spectra, and according to the measurements made by the GCMS⁹, features at $m/z = 40$ and $m/z = 44$ must be attributed respectively to instrument background ^{40}Ar and CO_2 . The features at $m/z = 28$ and 30 can be attributed mainly to $^{14}\text{N}^{14}\text{N}$ and $^{15}\text{N}^{15}\text{N}$, respectively. The fragment at $m/z = 17$ is enhanced (from $1,306$ to $7,000$ counts s^{-1}), which might be attributed to a greater amount of aerosols collected during the second ACP aerosol sampling. Furthermore, a potential pyrolysis fragment feature appears at $m/z = 27$, the intensity of which increases from 0 (background) to $1,100$ counts s^{-1} (analysis signal) on the spectra; on Fig. 1, this feature was obviously lost in the background. Although the NIST library molecular fragment list suggests that several compounds (C_2H_4 , C_2H_6 , C_3H_8 and HCN) may contribute to this feature at $m/z = 27$, a chain of deduction (see main text) allows us to conclude that HCN is the main contributor to the feature at $m/z = 27$. It is possible that there is instrumental crosstalk between adjacent features at $m/z = 27$ and 28 ; that is, some ions with $m/z = 28$ can be detected at $m/z = 27$, as observed by the GCMS experiment⁹. This crosstalk should depend on the intensity of the $m/z = 28$ signal and vary in proportion to it. In fact, we noticed a strongly varying ($m/z = 28$)/($m/z = 27$) ratio, which indicates that the crosstalk plays here a minor role (see Supplementary Fig. 5). An experimental confirmation is expected by performing the injection of HCN into the laboratory models of the ACP-GCMS experiments, identical to the models having flown on the Huygens probe.

tholins) of Titan's aerosols. These tholins have optical properties similar to those of Titan's aerosols¹⁶, and the simulation experiments that produce them also yield a suite of organic gases qualitatively representative of the hydrocarbon and nitrile gases observed in Titan's stratosphere²⁰.

A few important but limited pyrolysis-GCMS analyses of tholin material show the release of NH_3 and HCN ^{15,19,21}, but a more detailed and systematic laboratory investigation is clearly needed.

Some limitations of the ACP-GCMS data are apparent. For example, as found in recent laboratory experiments (see Supplementary Fig. 1), a feature at $m/z = 78$ is diagnostic of the main MS fragment of benzene, an aromatic compound expected to be present in Titan's aerosol according to certain chemical models^{4,12}. As the data are not sufficiently reliable above $m/z = 50$, such aromatics are not directly detectable. The best approach is to infer higher-mass products using quantitative analyses from laboratory experiments now in progress. In particular, the presence of NH_3 in the pyrolysis

products provides information about the general class of aerosols in Titan's atmosphere. Analytic diagnoses of some tholins have indicated the presence of the NH_2 chemical group, which should be the major contributor to NH_3 released by pyrolysis (see Supplementary Information, Section 3). Evidence supporting this idea has been obtained from recent mass spectrometric measurements²². The presence of HCN is less informative, as this molecule is generally observed in the thermal decomposition of polymeric species that include nitrile CN chemical groups or structural CN bonds^{23,24} (see Supplementary Information, Section 3). The presence of both NH_3 and HCN does demonstrate, however, that nitrogen can be incorporated into Titan's aerosol in different ways, and the general presence of nitrogen in the aerosol suggests that the aerosol acts as an important sink for atmospheric nitrogen²⁵.

Finally, our measurements do not indicate any substantial difference between the two samples collected at different altitudes, covering the range from the middle stratosphere down to the middle troposphere. This is consistent with an aerosol of homogeneous composition between altitudes of 130 and 20 km, suggesting a common source. Such a source is generally associated with a photochemical production layer well above 200 km (ref. 7). From this upper region of the atmosphere the aerosol is transported to lower levels by diffusion, precipitation and atmospheric circulation.

The results presented here show that the complex organic matter produced by Titan's atmospheric chemistry is being carried irreversibly to the surface by the aerosols. This material should thus contribute to the composition of the surface and to its spectral signatures, in particular in the infrared range. Other complementary information on the optical and radiometric properties of the aerosols can be obtained from the Descent Imager/Spectral Radiometer instrument²⁶ on the Huygens probe, and from the Visible and Infrared Mapping Spectrometer and the Composite Infrared Spectrometer on board the Cassini spacecraft.

Received 27 May; accepted 20 October 2005.

Published online 30 November 2005.

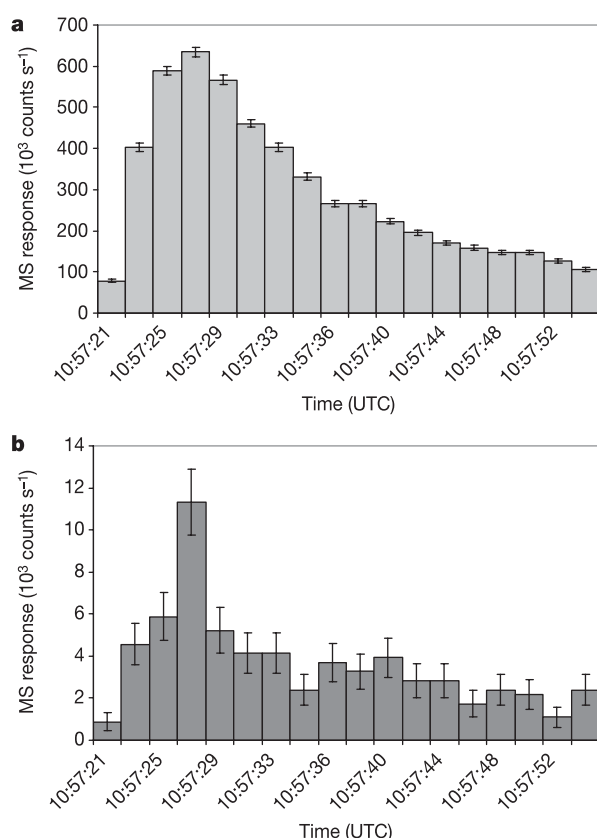


Figure 3 | Temporal evolution of the intensity of the MS signatures attributed to ammonia for the second aerosol sample pyrolysis. a, b, The temporal evolution of the signal for $m/z = 16$ (a) and $m/z = 17$ (b) during the transfer of the gases evolved from the second aerosol sample pyrolysis at 600 °C (named ACP 2-3). Error bars represent $\pm$ one standard deviation ($\pm 1\sigma$) to the experimental value measured. A signature at $m/z = 17$ is observed on the mass spectra measured during the transfer of the gases evolved from the first and the second aerosol samples pyrolysed at 600 °C (see Figs 1 and 2). The rate at which the signal of the feature builds during successive sweeps of the MS decays with time. This is characteristic of a real signal, and not of noise (if the event at $m/z = 17$ was due to noise, its intensity would not decrease with time during the successive injections in the MS). As seen on this figure, during the considered transfer, the signal at $m/z = 17$ evolves in the same way as the signal at $m/z = 16$ during ACP 2-3 transfer. However this behaviour cannot be explained by the contribution of the $^{13}\text{CH}_4$ alone. It requires the contribution of another species. Taking into account the distribution of the mass spectrum and the NIST MS library, the only possible candidate is NH_3 . A similar behaviour is observed on the mass spectra measured during the transfer of the gases evolved from the first aerosol sample pyrolysis at 600 °C.

- McKay, C. P., Pollack, J. B. & Courtin, R. The greenhouse and anti-greenhouse effects on Titan. *Science* **253**, 1118–1121 (1991).
- Rannou, P., Hourdin, F., McKay, C. P. & Luz, D. A coupled dynamics-microphysics model of Titan's atmosphere. *Icarus* **170**, 443–462 (2004).
- DelGenio, A. D., Zhou, W. & Eichler, T. P. Equatorial superrotation in a slowly rotating GCM — Implications for Titan and Venus. *Icarus* **101**, 1–17 (1993).
- Lebonnois, S., Bakes, E. L. O. & McKay, C. P. Transition from gaseous compounds to aerosols in Titan's atmosphere. *Icarus* **159**, 505–517 (2002).
- Barth, E. L. & Toon, O. W. Properties of methane clouds on Titan: results from microphysical modelling. *Geophys. Res. Lett.* **31**, L17S07 (2004).
- Mayo, L. A. & Samuelson, R. E. Condensate clouds in Titan's north polar stratosphere. *Icarus* **176**, 316–330 (2005).
- McKay, C. P. *et al.* Physical properties of the organic aerosols and clouds of Titan. *Planet. Space Sci.* **49**, 79–99 (2001).
- Niemann, H. *et al.* The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe. *Nature* doi:10.1038/nature04122 (this issue).
- Israel, G., Cabane, M., Raulin, F., Chassefière, E. & Boon, J. J. Aerosols in Titan's atmosphere: models, sampling techniques and chemical analysis. *Ann. Geophys.* **9**, 1–13 (1991).
- Israel, G. *et al.* Huygens Probe Aerosol Collector Pyrolyser Experiment. *Space Sci. Rev.* **104**, 435–466 (2002).
- Niemann, H. *et al.* The Gas Chromatograph Mass Spectrometer for the Huygens Probe. *Space Sci. Rev.* **104**, 551–590 (2002).
- Wilson, E. H. & Atreya, S. K. Chemical sources of haze formation in Titan's atmosphere. *Planet. Space Sci.* **51**, 1017–1033 (2003).
- Chang, S., DesMarais, D., Mack, R., Miller, S. L. & Strathearn, G. E. in *Earth's Earliest Biosphere: Its Origin and Evolution* (ed. Schopf, J. W.) 53–92 (Princeton Univ. Press, Princeton, 1982).
- Bar-Nun, A., Kleinfeld, I. & Ganor, E. Shape and optical properties of aerosols formed by photolysis of acetylene, ethylene and hydrogen cyanide. *J. Geophys. Res.* **93**, 8383–8387 (1988).
- Clarke, D. W. & Ferris, J. P. Titan haze: structure and properties of cyanoacetylene and cyanoacetylene-acetylene photopolymers. *Icarus* **127**, 158–172 (1997).
- Khare, B. N. *et al.* Optical constants of organic tholins produced in a simulated titanian atmosphere: from X-ray to microwave frequencies. *Icarus* **60**, 127–137 (1984).

17. Coll, P. *et al.* Experimental laboratory simulation of Titan's atmosphere: aerosols and gas phase. *Planet. Space Sci.* **47**, 1331–1340 (1999).
18. Imanaka, H. *et al.* Laboratory experiments of Titan tholins formed in cold plasma at various pressures: implications for nitrogen containing polycyclic aromatic compounds in Titan haze. *Icarus* **168**, 344–366 (2004).
19. Khare, B. N. *et al.* The organic aerosols of Titan. *Adv. Space Res.* **4**, 59–68 (1984).
20. Coll, P., Coscia, D., Gazeau, M.-C., Guez, L. & Raulin, F. Review and latest results of laboratory investigations of Titan's aerosols. *Orig. Life Evol. Biosph.* **28**, 195–213 (1997).
21. Ehrenfreund, P. *et al.* Analytical pyrolysis experiments of Titan aerosol analogues in preparation for the Cassini Huygens mission. *Adv. Space Res.* **15**, 335–342 (1995).
22. Somogyi, A., Oh, C.-H., Smith, M. A. & Lunine, J. I. Organic environments on Saturn's moon, Titan: simulating chemical reactions and analysing products by FT-ICR and ion trap mass spectrometry. *J. Am. Soc. Mass Spectrom.* **16**, 850–859 (2005).
23. Thompson, W. R., Henry, T. J., Schwartz, J. M., Khare, B. N. & Sagan, C. Plasma discharge in $N_2 + CH_4$ at low pressures: experimental results and applications to Titan. *Icarus* **90**, 57–73 (1991).
24. Minard, R. D., Hatcher, P. G., Gourley, R. C. & Matthews, C. N. Structural investigations of hydrogen cyanide polymers: new insights using TMAH thermochemolysis/GC-MS. *Orig. Life Evol. Biosph.* **28**, 461–473 (1998).
25. McKay, C. P. Elemental composition, solubility, and optical properties of Titan's organic haze. *Planet. Space Sci.* **44**, 741–747 (1996).
26. Tomasko, M. G. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* doi:10.1038/nature04126 (this issue).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge financial support from CNES, CNRS, the Austrian Ministry of Research and NASA. For the fabrication and qualification of ACP's mechanical and pneumatic components, the prime industrial contractor was SNECMA, and we acknowledge CNES Toulouse who strongly supported the CNRS ACP team. In particular, we thank C. Gelas, R. Salomé and E. Condé. We also thank the other main contractors, The Joanneum Research Institute and Austrian Aerospace, for the quality of their work. We acknowledge the support of the following people during the development of the instrument: M.-C. Gazeau and the LISA team at Créteil, R. Sablé and the ONERA/CERT in Toulouse, C. Cordelle and F. Marchandise at SNECMA, and G. Zeynard at Austrian Aerospace. We are also much indebted to the ESA Huygens Project Team for its constant support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.I. (Guy.israel@aerov.jussieu.fr).

LETTERS

The vertical profile of winds on Titan

M. K. Bird¹, M. Allison², S. W. Asmar³, D. H. Atkinson⁴, I. M. Avruch⁵, R. Dutta-Roy¹, Y. Dzierma¹, P. Edenhofer⁶, W. M. Folkner³, L. I. Gurvits⁵, D. V. Johnston³, D. Plettemeier⁷, S. V. Pogrebenko⁵, R. A. Preston³ & G. L. Tyler⁸

One of Titan's most intriguing attributes is its copious but featureless atmosphere. The Voyager 1 fly-by and occultation in 1980 provided the first radial survey of Titan's atmospheric pressure and temperature^{1,2} and evidence for the presence of strong zonal winds³. It was realized that the motion of an atmospheric probe could be used to study the winds, which led to the inclusion of the Doppler Wind Experiment⁴ on the Huygens probe⁵. Here we report a high resolution vertical profile of Titan's winds, with an estimated accuracy of better than 1 m s^{-1} . The zonal winds were prograde during most of the atmospheric descent, providing *in situ* confirmation of superrotation on Titan. A layer with surprisingly slow wind, where the velocity decreased to near zero, was detected at altitudes between 60 and 100 km. Generally weak winds ($\sim 1 \text{ m s}^{-1}$) were seen in the lowest 5 km of descent.

Titan's winds have been the subject of many investigations since that first close-up look from Voyager nearly 25 years ago. The infrared observations revealed a distinct pole-to-equator latitudinal contrast in temperature, varying from $\Delta T \approx 3 \text{ K}$ at the surface to $\Delta T \approx 20 \text{ K}$ in the stratosphere, implying a superrotational, global cyclostrophic circulation analogous to that observed on Venus³. Scaling for a hydrostatic, gradient-balanced flow suggested that the meridional and vertical winds should be much weaker than the zonal motion. Titan-specific general circulation models (GCMs) have since been introduced to study the conditions necessary for generation of atmospheric superrotation⁶⁻⁹.

Observational evidence for winds on Titan has also been inferred from the finite oblateness of surfaces of constant pressure determined from precise ground-based astrometry during stellar occultations in 1989 and 2001^{10,11}. These occultation experiments, as well as the thermal gradient observations, cannot be used to determine the sense of the zonal winds (that is, prograde or retrograde). A technique offering a direct determination of the wind velocity is to measure the differential Doppler shift of atmospheric spectral features as the field-of-view moves from east limb to west limb. Infrared heterodyne observations of Titan's ethane emission at $12 \mu\text{m}$ have yielded evidence for prograde winds with velocities exceeding 200 m s^{-1} but with a relatively large uncertainty of $\pm 150 \text{ m s}^{-1}$ (ref. 12). These results assume a global-average zonal wind field and apply to only a limited range in height near the 1 hPa level (200 km altitude). More traditional cloud-tracking techniques using Voyager 1 and ground-based images of Titan have been largely stymied by the ubiquitously poor image contrast. The success of such efforts has improved with the extended capabilities of the imaging system on Cassini, from which a number of atmospheric features have been identified as middle- to lower-tropospheric clouds, particularly near Titan's southern pole¹³.

The Huygens probe entered and descended for nearly 150 min through the atmosphere of Titan, survived impact on the surface, and continued its telemetry broadcast to the Cassini spacecraft on two

separate radio links, denoted channels A and B, for an additional 193 min (ref. 5). The Doppler Wind Experiment (DWE) instrumentation—consisting of an atomic rubidium oscillator in the probe transmitter to assure adequate frequency stability of the radiated signal and a similar device in the orbiter receiver to maintain the high frequency stability—was implemented only in channel A (2,040 MHz)⁴. Whereas channel B (2,098 MHz) functioned flawlessly during the entire mission, the channel A receiver was not properly configured during the probe relay sequence. All data on channel A, including the probe telemetry and the planned DWE measurements, were thus lost.

The channel A signal was monitored on Earth during the Huygens mission at fifteen radio telescopes, six of which recorded ground-based DWE measurements of the carrier frequency. Details on the

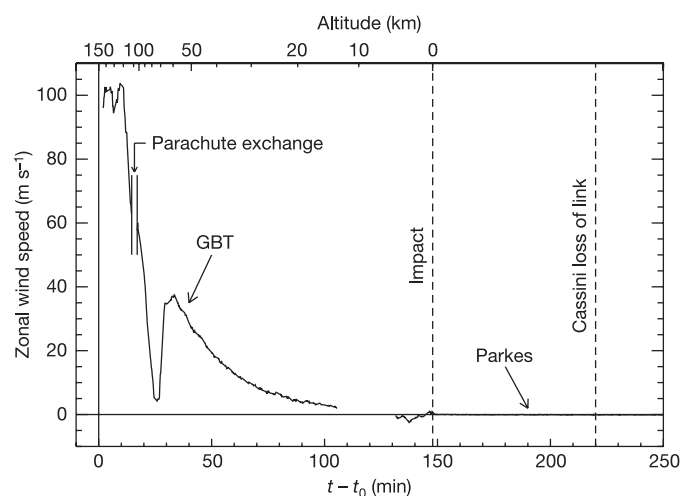


Figure 1 | Zonal wind velocity during the Huygens mission. The winds aloft are strictly prograde (positive zonal wind), but a significant reduction in the wind speed is observed at altitudes in the interval from 60 km to beyond 100 km. A one-minute interval associated with the parachute exchange, for which a more accurate determination of the actual descent velocity is necessary, has been excluded from this preliminary analysis. A monotonic decrease in the zonal wind speed is recorded from 60 km down to the end of the GBT (Green Bank Telescope) track at 10:56 SCET/UTC. The Parkes observations (from the Parkes Radio Telescope) could not begin until 11:22 SCET/UTC, thereby excluding wind determinations at heights from roughly 5 to 14 km. By this time Huygens was in a region of weak winds ($|U| \approx 1 \text{ m s}^{-1}$) that display distinct structure with a trend towards retrograde motion. The 26-min gap between GBT and Parkes may be closed later with additional Doppler measurements from participating radio telescopes located in the intervening longitude range, compare Supplementary Table 1. t_0 indicates the time of the start of descent.

¹Radioastronomisches Institut, Universität Bonn, Auf dem Hügel 71, 53125 Bonn, Germany. ²NASA Goddard Institute for Space Studies, 2880 Broadway, New York, New York 10025, USA. ³Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109, USA. ⁴Department of Electrical & Computer Engineering, University of Idaho, Moscow, Idaho 83844-1023, USA. ⁵Joint Institute for VLBI in Europe, PO Box 2, 7990 AA Dwingeloo, The Netherlands. ⁶Institut für HF-Technik, Universität Bochum, 44780 Bochum, Germany. ⁷Elektrotechnisches Institut, Technische Universität Dresden, 01062 Dresden, Germany. ⁸Center for Radar Astronomy, Stanford University, Stanford, California 94305, USA.

participants in the radio astronomy segment of the Huygens mission, the observation campaign, and plots of the raw data are given in Supplementary Information. Only the data sets from the NRAO Robert C. Byrd Green Bank Telescope (GBT) in West Virginia and the CSIRO Parkes Radio Telescope in Australia have been processed for this initial report.

Starting with the raw Doppler measurements, it is essentially a geometric exercise to derive the motion of the transmitter in the Titan frame of reference. Motion of the Huygens probe in the vertical direction is measured *in situ* by many different instruments, primarily pressure sensors¹⁴, but also by the Huygens radar altimeters. The consolidated measurements are processed iteratively by the Huygens Descent Trajectory Working Group (DTWG) to produce a series of continually improving probe trajectories referenced to Titan, including the descent velocity profile required for the present analysis⁵. The results presented here are based on DTWG data set no. 3, released in May 2005. The DTWG, based on knowledge of the Huygens trajectory to the point of atmospheric entry, also supplies the estimated spatial coordinates of the Huygens probe in latitude ($10.33 \pm 0.17^\circ$ S), longitude ($196.08 \pm 0.25^\circ$ W) and altitude (154.8 ± 11.2 km) at the start of descent, time t_0 .

Two key assumptions about the horizontal motion of the probe simplify the problem. The first of these is that the horizontal drift of the probe follows the horizontal wind with a negligible response time. The actual response time for the Huygens descent system is estimated to be roughly 30–40 s in the stratosphere, decreasing to 3–5 s in the lowest 10 km (ref. 15). It follows that this first assumption may be fulfilled only marginally during the early minutes of the descent and during the change to a smaller (stabilizer) parachute at $t = t_0 + 15$ min ('parachute exchange'). The second assumption, that the drift in the meridional (north–south) direction is negligible, is based on theoretical considerations that imply dominance of the zonal (east–west) atmospheric circulation^{3,6–9}. Under these conditions one is able to eliminate all other known contributions to the measured Doppler shift and determine the one remaining unknown, the zonal wind velocity. In addition to the above mentioned descent velocity, which slowly decreases with decreasing altitude, small, nearly constant, corrections must be applied for the effect of special relativity (-7.5 Hz, where the minus sign means a red shift), as well as the effects of general relativity associated with the Sun (18.2 Hz), Saturn (-0.7 Hz), Earth (1.4 Hz) and Titan (-0.08 Hz). Propagation corrections to the Doppler measurements from the neutral and ionized intervening media (Titan, interplanetary, Earth) have been estimated and found to be negligible. Finally, a small correction of $+10.0$ Hz was applied to the absolute transmission frequency by requiring that Huygens remain stationary on Titan's surface after landing. This residual is within the error limits of the pre-launch unit-level calibration of $+9.2$ Hz determined for the specific DWE rubidium oscillator unit used to drive the Huygens channel A transmitter.

The zonal wind derived from the ground-based Doppler data is shown in Fig. 1 as a function of time. More precisely, this quantity is the horizontal eastward velocity of Huygens with respect to the surface of Titan (with a positive value indicating the prograde direction). The time-integrated wind measurement from t_0 yields an estimate for the longitude of the Huygens landing site on Titan,

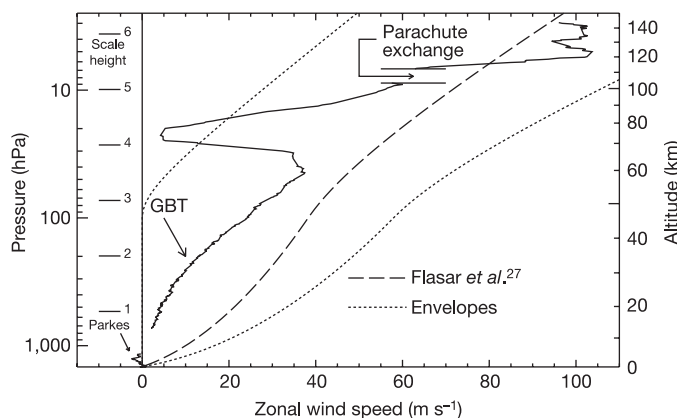


Figure 2 | Titan zonal wind height profile. The zonal wind derived from GBT and Parkes observations is compared with the prograde Titan engineering wind model and envelopes based on Voyager temperature data²⁷. The estimated uncertainties in the zonal wind speed, based on an adaptation of an error analysis for the Huygens-Cassini link²⁸, are of the order of 80 cm s^{-1} at high altitude and drop roughly in proportion to the absolute speed to 15 cm s^{-1} just above the surface. These uncertainties are primarily systematic errors associated with the Huygens trajectory at entry. The estimated statistical (measurement) error is always smaller, the standard deviation being of the order of $\sigma \approx 5 \text{ cm s}^{-1}$ towards the end of descent. With the possible exception of the region above 100 km, where the wind fluctuations are greatest, the zonal flow is found to be generally weaker than those of the model. The wind shear layer in the height range between 60 and beyond 100 km was unexpected and is at present unexplained.

$192.33 \pm 0.31^\circ$ W, which corresponds to an eastward drift of $3.75 \pm 0.06^\circ$ (165.8 ± 2.7 km) over the duration of the descent. Unfortunately, because of the slow rotation of Titan and the fact that the Earth was near zenith as viewed by Huygens, the Doppler data recorded after landing are not considered suitable for a more precise determination of the Huygens longitude.

The variation of the zonal wind with altitude and pressure level is shown in Fig. 2. The measured profile roughly agrees with the upper level wind speeds anticipated by the engineering model, and is generally prograde above 14 km altitude. Assuming this local observation is representative of conditions at this latitude, the large prograde wind speed measured between 45 and 70 km altitude and above 85 km is much larger than Titan's equatorial rotation speed ($\Omega a \approx 11.74 \text{ m s}^{-1}$, where $\Omega = 4.56 \times 10^{-6} \text{ rad s}^{-1}$ and $a = 2,575 \text{ km}$ are Titan's rotation rate and radius, respectively), and thus represents the first *in situ* confirmation of the inferred superrotation of the atmosphere at these levels, as anticipated from the Voyager temperature data³. Moreover, the measured winds are consistent with the strong winds inferred from ground-based data under the assumption of cyclostrophic balance^{10–12}.

The most striking departure of the measured profile from the engineering model is the region of strong reversed shear between 65 and 75 km altitude (approximately 40 and 25 hPa, respectively), where the speed decreases to a minimum of 4 m s^{-1} , which then reverts to strong prograde shear above 75 km. This feature of Titan's wind profile is unlike that measured by any of the Doppler-tracked probes in the atmosphere of Venus¹⁶.

Table 1 | Predictions of Titan's meteorology and DWE results

Prediction/model feature	DWE result	References
Atmospheric superrotation ($U \gg 12 \text{ m s}^{-1}$) [*]	Verified for upper troposphere and stratosphere	3, 6–8
Prograde (eastward) flow ($U > 0$)	Verified for all levels above 15 km (GBT data)	6–8, 23, 24
Isolated reversed shear ($\partial U / \partial z < 0$) within lower stratosphere	Verified, but stronger than simulated at 65–75 km, with $Ri \approx 2$	7–9
Geostrophic ($U \ll 12 \text{ m s}^{-1}$) sub-layer near surface	Verified and deeper than anticipated (more than ~ 1 scale height)	7–9, 25
Very weak surface winds	Verified ($ U \approx 1 \text{ m s}^{-1}$)	23, 25
Warmer-poleward near-surface temperature in southern hemisphere	Consistent with geostrophic balance of upward-westward shear of low-level winds	18, 26

^{*} U = zonal wind velocity.

The preliminary wind data shown in Fig. 2 have provided an *in situ* test of several Titan weather predictions, as summarized in Table 1. The verification of a superrotating atmosphere (in which the zonal wind velocity U is faster than the solid surface beneath it) definitively places Titan's meteorology in the same regime as that of Venus. The confirmed prograde direction of the flow lends further evidence for dynamical control of the cyclostrophic thermal structure. The isolated, reverse vertical shear region in the lower stratosphere, while not expected by the Huygens science team, appears to be present as a similar, but weaker, structure in Titan GCM simulations of this region^{7–9}. We estimate that the implied Richardson number is as small as $Ri \approx 2$ –5 near the 30 hPa level. Earlier studies have shown that potential vorticity mixing within atmospheric regions where $Ri \approx 2$ imposes a relatively flat variation of wind velocity over latitude, as compared with the rapid poleward increase of wind wherever Ri is large¹⁷.

Although models and theory have anticipated a geostrophic sub-layer, where the atmospheric flow is much slower than Titan's surface rotation speed, it is interesting that this appears to extend more than one scale height above the surface. This contrasts with the engineering wind model, which suggests a stronger vertical shear near the surface, and raises the interesting possibility of a more Earth-like weather regime within Titan's lower troposphere, perhaps including an alternation of low- and high-pressure centres and some meridional motion there. Surface winds are measured to be weak ($|U| \approx 1 \text{ m s}^{-1}$), as expected. The vertical wind shear in the lowest part of the troposphere, with increasing westward flow with altitude, implies (via the thermal wind equation) a geostrophically balanced warmer-poleward temperature structure near the surface. Although some caution must be applied to a global interpretation of the available single-latitude measurement, this feature of the DWE profile would be consistent with the relatively warm southern (summer) pole, which is considered to be the underlying reason for the convective cloud features observed there¹⁸.

It will be of great interest to see if these inferences from comparisons between the models and the measured DWE profile are corroborated by the vertical-latitudinal sounding of temperatures within the 10–50 hPa region by the Cassini Composite Infrared Spectrometer (CIRS) instrument^{19,20}—or by the pressure/temperature versus height profiles down to the surface derived from Cassini radio science measurements²¹ during the upcoming Titan occultations.

The ambiguity between contributions from zonal and meridional winds, at least near the surface, has essentially been resolved by a detailed comparison with wind drift data from the Descent Imager/Spectral Radiometer (DISR) instrument²². The current DISR analysis indicates that Huygens drifted roughly westward during the last 7 km of descent. This is consistent with the small, but predominantly retrograde, zonal wind determinations from the Parkes Doppler data recorded during the 15 min before landing on Titan.

As seen in Fig. 2, there is a data gap of less than a half scale-height between the region of smooth eastward flow above 14 km and the apparently more structured, but very weak, wind regime ($U < 2 \text{ m s}^{-1}$) from 5 km down to the surface. A future report should be able to address this transition on the basis of further Doppler data from the other ground-based radio telescopes (see Supplementary Information). In addition, the simultaneously recorded Very Long Baseline Interferometry (VLBI) measurements of the position of the probe on the sky should eventually allow the assumption of purely zonal flow to be dropped. A combined Doppler/VLBI solution would then yield the full two-dimensional horizontal wind profile during the Huygens descent.

Received 20 May; accepted 20 July 2005.

Published online 30 November 2005.

1. Tyler, G. L. *et al.* Radio science investigations of the Saturn system with Voyager 1: Preliminary results. *Science* **212**, 201–206 (1981).

2. Lindal, G. F. *et al.* The atmosphere of Titan: An analysis of the Voyager 1 radio occultation measurements. *Icarus* **53**, 348–363 (1983).
3. Flasar, F. M., Samuelson, R. E. & Conrath, B. J. Titan's atmosphere: temperature and dynamics. *Nature* **292**, 693–698 (1981).
4. Bird, M. K. *et al.* The Huygens Doppler Wind Experiment. *Space Sci. Rev.* **104**, 613–640 (2002).
5. Lebreton, J.-P. *et al.* An overview of the descent and landing of the Huygens probe. *Nature* doi:10.1038/nature04347 (this issue).
6. Del Genio, A. D., Zhou, W. & Eichler, T. P. Equatorial superrotation in a slowly rotating GCM: Implications for Titan and Venus. *Icarus* **101**, 1–17 (1993).
7. Hourdin, F. *et al.* Numerical simulation of the general circulation of the atmosphere of Titan. *Icarus* **117**, 358–374 (1995).
8. Tokano, T., Neubauer, F. M., Laube, M. & McKay, C. P. Seasonal variation of Titan's atmospheric structure simulated by a general circulation model. *Planet. Space Sci.* **47**, 493–520 (1999).
9. Luz, D., Hourdin, F., Rannou, P. & Lebonnois, S. Latitudinal transport by baroclinic waves in Titan's stratosphere. II. Results from a coupled dynamics-microphysics-photochemistry GCM. *Icarus* **166**, 343–358 (2003).
10. Hubbard, W. B. *et al.* The occultation of 28 Sgr by Titan. *Astron. Astrophys.* **269**, 541–563 (1993).
11. Bouchez, A. H. *Seasonal Trends in Titan's Atmosphere: Haze, Wind and Clouds* PhD thesis, California Inst. Technol. (2003); available at (<http://www.gps.caltech.edu/~antonin/thesis>).
12. Kostiuik, T. *et al.* Direct measurement of winds on Titan. *Geophys. Res. Lett.* **28**, 2361–2364 (2000).
13. Porco, C. C. *et al.* Imaging of Titan from the Cassini spacecraft. *Nature* **434**, 159–168 (2005).
14. Fulchignoni, M. *et al.* *In situ* measurements of the physical characteristics of Titan's atmosphere. *Nature* doi:10.1038/nature04314 (this issue).
15. Atkinson, D. H., Pollack, J. B. & Seiff, A. Measurement of a zonal wind profile on Titan by Doppler tracking of the Cassini entry probe. *Radio Sci.* **25**, 865–882 (1990).
16. Counselman, C. C. III, Gourevitch, S. A., King, R. W. & Lioriot, G. B. Zonal and meridional circulation of the lower atmosphere of Venus determined by radio interferometry. *J. Geophys. Res.* **85**, 8026–8030 (1980).
17. Allison, M., Del Genio, A. D. & Zhou, W. Zero potential vorticity envelopes for the zonal-mean velocity of the Venus/Titan atmospheres. *J. Atmos. Sci.* **51**, 694–702 (1994).
18. Brown, M. E., Bouchez, A. H. & Griffith, C. A. Direct detection of variable tropospheric clouds near Titan's south pole. *Nature* **420**, 795–797 (2002).
19. Flasar, F. M. *et al.* Exploring the Saturn system in the thermal infrared: The Composite Infrared Spectrometer. *Space Sci. Rev.* **115**, 169–297 (2004).
20. Flasar, F. M. *et al.* Titan's atmospheric temperatures, winds, and composition. *Science* **308**, 975–978 (2005).
21. Kliore, A. J. *et al.* Cassini radio science. *Space Sci. Rev.* **115**, 1–70 (2004).
22. Tomasko, M. G. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* doi:10.1038/nature04126 (this issue).
23. Flasar, F. M. & Conrath, B. J. in *Proc. Symposium on Titan* (ed. Kaldeich, B.) 89–99 (ESA-SP 338, ESA Publications, ESTEC, Noordwijk, The Netherlands, 1992).
24. Del Genio, A. D. & Zhou, W. Simulations of superrotation on slowly rotating planets: Sensitivity to rotation and initial condition. *Icarus* **120**, 332–343 (1996).
25. Allison, M. in *Proc. Symposium on Titan* (ed. Kaldeich, B.) 113–118 (ESA SP-338, ESA Publications, ESTEC, Noordwijk, The Netherlands, 1992).
26. Tokano, T. Meteorological assessment of the surface temperatures on Titan: Constraints on the surface type. *Icarus* **173**, 222–242 (2005).
27. Flasar, F. M., Allison, M. & Lunine, J. I. in *Huygens Science Payload and Mission* (ed. Wilson, A.) 287–298 (ESA-SP 1177, ESA Publications, ESTEC, Noordwijk, The Netherlands, 1997).
28. Dutta-Roy, R. & Bird, M. K. in *Planetary Probe Atmospheric Entry and Descent Trajectory Analysis and Science* (ed. Wilson, A.) 109–116 (ESA SP-544, ESA Publications, ESTEC, Noordwijk, The Netherlands, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This Letter presents results of a research project partially funded by the Deutsches Zentrum für Luft- und Raumfahrt (DLR). Parts of the research described here were carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA, and by NASA's Goddard Institute for Space Studies. We thank R. Kohl, K.-P. Wagner and M. Heyl for their efforts during the DWE development programme. We appreciate the support provided by the National Radio Astronomy Observatory (NRAO) and the Australia Telescope National Facility (ATNF). NRAO is operated by Associated Universities, Inc., under a cooperative agreement with the NSF. The ATNF, managed by the Commonwealth Scientific and Industrial Research Organization (CSIRO), is funded by the Commonwealth of Australia.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.K.B. (mbird@astro.uni-bonn.de).

Genome sequence, comparative analysis and haplotype structure of the domestic dog

Kerstin Lindblad-Toh¹, Claire M Wade^{1,2}, Tarjei S. Mikkelsen^{1,3}, Elinor K. Karlsson^{1,4}, David B. Jaffe¹, Michael Kamal¹, Michele Clamp¹, Jean L. Chang¹, Edward J. Kulbokas III¹, Michael C. Zody¹, Evan Mauceli¹, Xiaohui Xie¹, Matthew Breen⁵, Robert K. Wayne⁶, Elaine A. Ostrander⁷, Chris P. Ponting⁸, Francis Galibert⁹, Douglas R. Smith¹⁰, Pieter J. deJong¹¹, Ewen Kirkness¹², Pablo Alvarez¹, Tara Biagi¹, William Brockman¹, Jonathan Butler¹, Chee-Wye Chin¹, April Cook¹, James Cuff¹, Mark J. Daly^{1,2}, David DeCaprio¹, Sante Gnerre¹, Manfred Grabherr¹, Manolis Kellis^{1,13}, Michael Kleber¹, Carolyn Bardeleben⁶, Leo Goodstadt⁸, Andreas Heger⁸, Christophe Hitte⁹, Lisa Kim⁷, Klaus-Peter Koepfli⁶, Heidi G. Parker⁷, John P. Pollinger⁶, Stephen M. J. Searle¹⁴, Nathan B. Sutter⁷, Rachael Thomas⁵, Caleb Webber⁸, Broad Institute Genome Sequencing Platform* & Eric S. Lander^{1,15}

Here we report a high-quality draft genome sequence of the domestic dog (*Canis familiaris*), together with a dense map of single nucleotide polymorphisms (SNPs) across breeds. The dog is of particular interest because it provides important evolutionary information and because existing breeds show great phenotypic diversity for morphological, physiological and behavioural traits. We use sequence comparison with the primate and rodent lineages to shed light on the structure and evolution of genomes and genes. Notably, the majority of the most highly conserved non-coding sequences in mammalian genomes are clustered near a small subset of genes with important roles in development. Analysis of SNPs reveals long-range haplotypes across the entire dog genome, and defines the nature of genetic diversity within and across breeds. The current SNP map now makes it possible for genome-wide association studies to identify genes responsible for diseases and traits, with important consequences for human and companion animal health.

Man's best friend, *Canis familiaris*, occupies a special niche in genomics. The unique breeding history of the domestic dog provides an unparalleled opportunity to explore the genetic basis of disease susceptibility, morphological variation and behavioural traits. The position of the dog within the mammalian evolutionary tree also makes it an important guide for comparative analysis of the human genome.

The history of the domestic dog traces back at least 15,000 years, and possibly as far back as 100,000 years, to its original domestication from the grey wolf in East Asia^{1–4}. Dogs evolved through a mutually beneficial relationship with humans, sharing living space and food sources. In recent centuries, humans have selectively bred dogs that excel at herding, hunting and obedience, and in this process have created breeds rich in behaviours that both mimic human behaviours and support our needs. Dogs have also been bred for desired physical characteristics such as size, skull shape, coat colour and texture⁵,

producing breeds with closely delineated morphologies. This evolutionary experiment has produced diverse domestic species, harbouring more morphological diversity than exists within the remainder of the family Canidae⁶.

As a consequence of these stringent breeding programmes and periodic population bottlenecks (for example, during the World Wars), many of the ~400 modern dog breeds also show a high prevalence of specific diseases, including cancers, blindness, heart disease, cataracts, epilepsy, hip dysplasia and deafness^{7,8}. Most of these diseases are also commonly seen in the human population, and clinical manifestations in the two species are often similar⁹. The high prevalence of specific diseases within certain breeds suggests that a limited number of loci underlie each disease, making their genetic dissection potentially more tractable in dogs than in humans¹⁰.

Genetic analysis of traits in dogs is enhanced by the close relationship between humans and canines in modern society.

¹Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, Massachusetts 02141, USA. ²Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. ³Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁴Program in Bioinformatics, Boston University, 44 Cummings Street, Boston, Massachusetts 02215, USA. ⁵Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, 4700 Hillsborough Street, Raleigh, North Carolina 27606, USA. ⁶Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095, USA. ⁷National Human Genome Research Institute, National Institutes of Health, 50 South Drive, MSC 8000, Building 50, Bethesda, Maryland 20892-8000, USA. ⁸MRC Functional Genetics, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK. ⁹UMR 6061 Genetique et Developpement, CNRS—Université de Rennes 1, Faculté de Médecine, 2, Avenue Leon Bernard, 35043 Rennes Cedex, France. ¹⁰Agencourt Bioscience Corporation, 500 Cummings Center, Suite 2450, Beverly, Massachusetts 01915, USA. ¹¹Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, California 94609, USA. ¹²The Institute for Genomic Research, Rockville, Maryland 20850, USA. ¹³Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. ¹⁴The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ¹⁵Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA.

*A list of participants and affiliations appears at the end of the paper.

Through the efforts of the American Kennel Club (AKC) and similar organizations worldwide, extensive genealogies are easily accessible for most purebred dogs. With the exception of human, dog is the most intensely studied animal in medical practice, with detailed family history and pathology data often available⁸. Using genetic resources developed over the past 15 years^{11–16}, researchers have already identified mutations in genes underlying ~25 mendelian diseases^{17,18}. There are also growing efforts to understand the genetic basis of phenotypic variation such as skeletal morphology^{10,19}.

The dog is similarly important for the comparative analysis of mammalian genome biology and evolution. The four mammalian genomes that have been intensely analysed to date (human^{20–22}, chimpanzee²³, mouse²⁴ and rat²⁵) represent only one clade (Euarchontoglires) out of the four clades of placental mammals. The dog represents the neighbouring clade, Laurasiatheria²⁶. It thus serves as an outgroup to the Euarchontoglires and increases the total branch length of the current tree of fully sequenced mammalian genomes, thereby providing additional statistical power to search for conserved functional elements in the human genome^{24,27–33}. It also helps us to draw inferences about the common ancestor of the two clades, called the boreoeutherian ancestor, and provides a bridge to the two remaining clades (Afrotheria and Xenarthra) that should be helpful for anchoring low-coverage genome sequence currently being produced from species such as elephant and armadillo²⁸.

Here we report a high-quality draft sequence of the dog genome covering ~99% of the euchromatic genome. The completeness, nucleotide accuracy, sequence continuity and long-range connectivity are extremely high, exceeding the values calculated for the recent draft sequence of the mouse genome²⁴ and reflecting improved algorithms, higher-quality data, deeper coverage and intrinsic genome properties. We have also created a tool for the formal assessment of assembly accuracy, and estimate that >99% of the draft sequence is correctly assembled.

We also report an initial compendium of SNPs for the dog population, containing >2.5 million SNPs derived primarily from partial sequence comparison of 11 dog breeds to a reference sequence. We characterized the polymorphism rate of the SNPs across breeds and the long-range linkage disequilibrium (LD) of the SNPs within and across breeds.

We have analysed these data to study genome structure, gene evolution, haplotype structure and phylogenetics of the dog. Our key findings include:

- The evolutionary forces molding the mammalian genome differ among lineages, with the average transposon insertion rate being lowest in dog, the deletion rate being highest in mouse and the nucleotide substitution rate being lowest in human.
- Comparison between human and dog shows that ~5.3% of the human genome contains functional elements that have been under purifying selection in both lineages. Nearly all of these elements are confined to regions that have been retained in mouse, indicating that they represent a common set of functional elements across mammals.
- Fifty per cent of the most highly conserved non-coding sequence in the genome shows striking clustering in ~200 gene-poor regions, most of which contain genes with key roles in establishing or maintaining cellular identity, such as transcription factors or axon guidance receptors.
- Sets of functionally related genes show highly similar patterns of evolution in the human and dog lineages. This suggests that we should be careful about interpreting accelerated evolution in human relative to mouse as representing human-specific innovations (for example, in genes involved in brain development), because comparable acceleration is often seen in the dog lineage.
- Analysis across the entire genome of the sequenced boxer and across 6% of the genome in ten additional breeds shows that linkage disequilibrium (LD) within breeds extends over distances of several megabases, but LD across breeds only extends over tens of kilobases.

These LD patterns reflect two principal bottlenecks in dog history: early domestication and recent breed creation.

- Haplotypes within breeds extend over long distances, with ~3–5 alleles at each locus. Portions of these haplotypes, as large as 100 kilobases (kb), are shared across multiple breeds, although they are present at widely varying frequencies. The haplotype structure suggests that genetic risk factors may be shared across breeds.
- The current SNP map has sufficient density and an adequate within-breed polymorphism rate (~1/900 base pairs (bp) between breeds and ~1/1,500 bp within breeds) to enable systematic association studies to map genes affecting traits of interest. Genotyping of ~10,000 SNPs should suffice for most purposes.
- The genome sequence can be used to select a small collection of rapidly evolving sequences, which allows nearly complete resolution of the evolutionary tree of nearly all living species of Canidae.

Generating a draft genome sequence

We sequenced the genome of a female boxer using the whole-genome shotgun (WGS) approach^{22,24} (see Methods and Supplementary Table S1). A total of 31.5 million sequence reads, providing ~7.5-fold sequence redundancy, were assembled with an improved version of the ARACHNE program³⁴, resulting in an initial assembly (CanFam1.0) used for much of the analysis below, and an updated assembly (CanFam2.0) containing minor improvements (Table 1 and Supplementary Table S2).

Genome assembly. The recent genome assembly spans a total distance of 2.41 Gb, consisting of 2.38 Gb of nucleotide sequence with the remaining 1% in captured gaps. The assembly has extremely high continuity. The N50 contig size is 180 kb (that is, half of all bases reside in a contiguous sequence of 180 kb or more) and the N50 supercontig size is 45.0 Mb (Table 1). In particular, this means that most genes should contain no sequence gaps and that most canine chromosomes (mean size 61 Mb) have nearly all of their sequence ordered and oriented within one or two supercontigs (Supplementary Table S2). Notably, the sequence contigs are ~50-fold larger than the earlier survey sequence of the standard poodle¹⁶.

The assembly was anchored to the canine chromosomes using data from both radiation hybrid and cytogenetic maps^{11,13,14}. Roughly 97% of the assembled sequence was ordered and oriented on the chromosomes, showing an excellent agreement with the two maps. There were only three discrepancies, which were resolved by obtaining additional fluorescence *in situ* hybridization (FISH) data from the sequenced boxer. The 3% of the assembly that could not be anchored consists largely of highly repetitive sequence, including eight supercontigs of 0.5–1.0 Mb composed almost entirely of satellite sequence.

The nucleotide accuracy and genome coverage of the assembly is high (Supplementary Table S3). Of the bases in the assembly, 98% have quality scores exceeding 40, corresponding to an error rate of less than 10^{-4} and comparable to the standard for the finished human sequence³⁵. When we directly compared the assembly to 760 kb of finished sequence (in regions where the boxer is

Table 1 | Assembly statistics for CanFam1.0 and 2.0

	CanFam1.0	CanFam2.0
N50 contig size	123 kb	180 kb
N50 supercontig size	41.2 Mb	45.0 Mb
Assembly size (total bases)	2.360 Gb	2.385 Gb
Number of anchored supercontigs	86	87
Percentage of genome in anchored supercontigs	96	97
Sequence in anchored bases	2.290 Gb	2.309 Gb
Percentage of assembly in gaps	0.9	0.8
Estimated genome size*	2.411 Gb	2.445 Gb
Percentage of assembly in 'certified regions', without assembly inconsistency	99.3	99.6

* Includes anchored bases, spanned gaps (21 Mb in CanFam1.0, 18 Mb in CanFam2.0) and centromeric sequence (3 Mb for each chromosome).

homozygous, to eliminate differences attributable to polymorphisms; see below), we found that the draft genome sequence covers 99.8% of the finished sequence and that bases with quality scores exceeding 40 have an empirical error rate of 2×10^{-5} (Supplementary Table S3).

Explaining the high sequence continuity. The dog genome assembly has superior sequence continuity (180 kb) than the WGS assembly of the mouse genome (25 kb) obtained several years ago²⁴. At least three factors contribute to the higher connectivity of the dog assembly (see Supplementary Information). First, we used a new version of ARACHNE with improved algorithms. Assembling the dog genome with the previous software version decreased N50 contig size from 180 kb to 61 kb, and assembling the mouse genome with the new version increased N50 contig size from 25 kb to 35 kb. Second, the amount of recently duplicated sequence is roughly twofold lower in dog than mouse (Supplementary Table S4); this improves contiguity because sequence gaps in both organisms tend to occur in recently duplicated sequence. Third, the dog sequence data has both higher redundancy (7.5-fold versus 6.5-fold) and higher quality (in terms of read length, pairing rate and tight distribution of insert sizes) compared with mouse. The contig size for the dog genome drops by about 32% when the data redundancy is decreased from 7.5-fold to 6.5-fold. A countervailing influence is that the dog genome contains polymorphism, whereas the laboratory mouse is completely inbred.

Assembly certification. Although ‘quality scores’ have been developed to indicate the nucleotide accuracy of a draft genome sequence³⁶, no analogous measures have been developed to reflect the long-range assembly accuracy. We therefore sought to develop such a measure on the basis of two types of internal inconsistencies (see Supplementary Information). The first is haplotype inconsistency, involving clear evidence of three or more distinct haplotypes within an assembled region from a single diploid individual. The second is linkage inconsistency, involving a cluster of reads for which the placement of the paired-end reads is illogical. This includes cases in which: (1) one end cannot be mapped to the region, (2) the linkage relationships are inconsistent with the sequence within contigs, or (3) distance constraints imply overlap between non-overlapping sequence contigs. The linkage inconsistency tests are most powerful when read pairs are derived from clone libraries with tight constraints on insert size. A region of assembly is defined as ‘certified’ if it is free of inconsistencies, and is otherwise ‘questionable’.

Approximately 99.6% of the assembly resides in certified regions, with the N50 size of certified regions being ~ 12 Mb or about one-fifth of a chromosome. The remaining questionable regions are typically small (most are less than 40 kb), although there are a handful of regions of several hundred kilobases (Supplementary Fig. S1 and Supplementary Tables S5, S6). The questionable regions typically contain many inconsistencies, probably reflecting mis-assembly or overcollapse owing to segmental duplication. Chromosomes 2, 11 and 16 have 1.0–2.0% of their sequence in questionable regions. The certified and questionable regions are annotated in the public release of the dog genome assembly. With the concept of assembly certification, the scientific community can have appropriate levels of confidence in the draft genome sequence.

Genome landscape and evolution

Our understanding of the evolutionary processes that shape mammalian genomes has greatly benefited from the comparative analysis of sequenced primate^{21,23} and rodent^{24,25} genomes. However, the rodent genome is highly derived relative to that of the common ancestor of the eutherian mammals. As the first extensive sequence from an outgroup to the clade that includes primates and rodents, the dog genome offers a fresh perspective on mammalian genome evolution. Accordingly, we examined the rates and correlations of large-scale rearrangement, transposon insertion, deletion and nucleotide divergence across three major mammalian orders (primates, rodents and carnivores).

Conserved synteny and large-scale rearrangements. We created multi-species synteny maps from anchors of unique, unambiguously aligned sequences (see Supplementary Information), showing regions of conserved synteny among dog, human, mouse and rat genomes. Approximately 94% of the dog genome lies in regions of conserved synteny with the three other species (Supplementary Figs S2–S4 and Supplementary Table S7).

Given a pair of genomes, we refer to a 'syntenic segment' as a region that runs continuously without alterations of order and orientation, and a 'syntenic block' as a region that is contiguous in two genomes but may have undergone internal rearrangements. Syntenic breakpoints between blocks reflect primarily interchromosomal exchanges, and breakpoints between syntenic segments reflect intrachromosomal rearrangements. In the analysis below, we focus on syntenic segments of at least 500 kb.

We identified a total of 391 syntenic breakpoints across dog, human, mouse and rat genomes (Fig. 1 and Supplementary Figs S2, S5). With data for multiple species, it is possible to assign events to specific lineages (Fig. 1 and Supplementary Table S8). We counted the total number of breakpoints along the human, dog, mouse and rat lineages, with the values for each rodent lineage reflecting all breakpoints since the common ancestor with human (Fig. 1). The total number of breakpoints in the human lineage is substantially smaller than in the dog, mouse or rat lineages (83 versus 100, 161 or 176, respectively). However, there are more intra-chromosomal breakpoints in the human lineage than in dog (52 versus 33).

Although the overall level of genomic rearrangement has been much higher in rodent than in human, comparison with dog shows that there are regions where the opposite is true. In particular, of the many intrachromosomal rearrangements previously observed between human chromosome 17 and the orthologous mouse

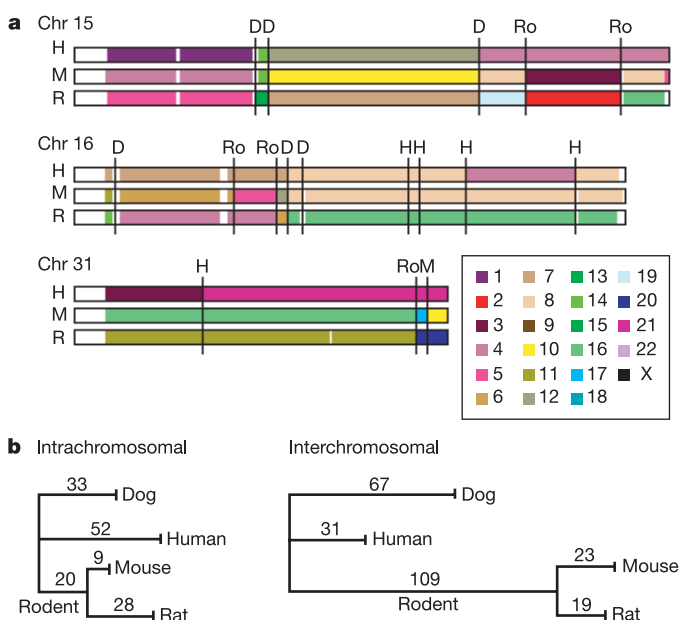


Figure 1 | Conserved synteny among the human, dog, mouse and rat genomes. **a**, Diagram of syntenic blocks (>500 kb) along dog chromosomes (Chr) 15, 16 and 31, with colours indicating the chromosome containing the syntenic region in other species. Synteny breakpoints were assigned to one of five lineages: dog (D), human (H), mouse (M), rat (R) or the common rodent ancestor (Ro). **b**, Lineage-specific intrachromosomal and interchromosomal breaks displayed on phylogenetic trees. Intrachromosomal breaks are seen more frequently in the human lineage than in mouse and rat, whereas interchromosomal breaks are somewhat more common in dog and considerably more common in rodents than in humans.

sequence²⁴, most have occurred in the human lineage (see Supplementary Information). Human chromosome 17 is rich in segmental duplications and gene families²¹, which may contribute to its genomic fragility^{37,38}.

Genomic insertion and deletion. The euchromatic genome of the dog is ~150 Mb smaller than in mouse, and ~500 Mb smaller than in human. The smaller total size is reflected at the local level, with 100-kb blocks of conserved synteny in dog corresponding to regions for which the median size is ~3% larger in mouse and ~15% larger in human.

To understand the balance of forces that determine genome size, we studied the alignments of the human, mouse and dog genomes (Fig. 2). In particular, we identified the lineage-specific interspersed repeats within each genome, which consist of particular families of short interspersed elements (SINEs), long interspersed elements (LINEs) and other transposable elements that are readily recognized by sequence analysis (Supplementary Tables S9, S10). The remaining sequence was annotated as 'ancestral', consisting of both ancestral unique sequence and ancestral repeat sequence; these two categories were combined because the power to recognize ancient transposon-derived sequences degrades with repeat age, particularly in the rapidly diverging mouse lineage²⁴.

This comparative analysis indicates that different forces account for the smaller genome sizes in dog and mouse relative to human. The smaller size of the dog genome is primarily due to the presence of substantially less lineage-specific repeat sequence in dog (334 Mb) than in human (609 Mb) or mouse (954 Mb). This reflects a lower activity of endogenous retroviral and DNA transposons (~26,000 extant copies in dog versus ~183,000 in human), as well as the fact that the SINE element in dog is smaller than in human (although of similar length to that in mouse). As a consequence, the total proportion of repetitive elements (both lineage-specific and ancestral) recognizable in the genome is lower for dog (34%) than for mouse (40%) or human (46%). In contrast, the smaller size of the mouse genome is primarily due to a higher deletion rate. Specifically, the amount of extant 'ancestral sequence' is much lower in mouse (1,474 Mb) than in human (2,216 Mb) or dog (1,997 Mb). Assuming an ancestral genome size of 2.8 Gb (ref. 24) and also that deletions occur continuously, we suggest that the rate of genomic deletion in the rodent lineage has been approximately 2.5-fold higher than in the

dog and human lineages (see Supplementary Information). As a consequence, the human genome shares ~650 Mb more ancestral sequence with dog than with mouse, despite our more recent common ancestor with the latter.

Active SINE family. Despite its relatively low proportion of transposable element-derived sequence, the dog genome contains a highly active carnivore-specific SINE family (defined as SINEC_Cf; RepBase release 7.11)¹⁶. The element is so active that many insertion sites are still segregating polymorphisms that have not yet reached fixation. Of ~87,000 young SINEC_Cf elements (defined by low divergence from the consensus sequence), nearly 8% are heterozygous within the draft genome sequence of the boxer. Moreover, comparison of the boxer and standard poodle genome sequences reveals more than 10,000 insertion sites that are bimorphic, with thousands more certain to be segregating in the dog population^{16,39}. In contrast, the number of polymorphic SINE insertions in the human genome is estimated to be fewer than 1,000 (ref. 40).

The biological effect of these segregating SINE insertions is unknown. SINE insertions can be mutagenic through direct disruption of coding regions or through indirect effects on regulation and processing of messenger RNAs³⁹. Such SINE insertions have already been shown to be responsible for two diseases in dog: narcolepsy and centronuclear myopathy^{41,42}. It is conceivable that the genetic variation resulting from these segregating SINE elements has provided important raw material for the selective breeding programmes that have produced the wide phenotypic variations among modern dog breeds^{16,43}.

Sequence composition. The human and mouse genomes differ markedly in sequence composition, with the human genome having slightly lower average G+C content (41% versus 42% in mouse) but much greater variation across the genome. The dog genome closely resembles the human genome in its distribution of G+C content (Fig. 3a; Spearman's $\rho = 0.85$ for dog-human and 0.76 for dog-mouse comparisons), even if we consider only nucleotides that can be aligned across all three species (Supplementary Fig. S6). The wider distribution of G+C content in human and dog is thus likely to reflect the boreoeutherian ancestor^{44,45}, with the more homogeneous composition in rodents having arisen primarily through lineage-specific changes in substitution patterns^{46,47} rather than deletion of sequences with high G+C content.

Rate of nucleotide divergence. We estimated the mean nucleotide divergence rates in 1-Mb windows along the dog, human and mouse lineages on the basis of alignments of all ancestral repeats, using the consensus sequence for the repeats as a surrogate outgroup (Fig. 3b; see also Supplementary Information).

The dog lineage has diverged more rapidly than the human lineage (median relative divergence rate of 1.18, longer branch length in 95% of windows), but at only half the rate of the mouse lineage (median relative rate of 0.48, shorter branch length in 100% of windows). The absolute divergence rates are somewhat sensitive to the evolutionary model used and the filtering of alignment artefacts (data not shown), but the relative rates appear to be robust and are consistent with estimates from smaller sequence samples with multiple outgroups^{28,48,49}. The lineage-specific divergence rates (human < dog < mouse) are probably explained by differences in metabolic rates^{50,51} or generation times^{52,53}, but the relative contributions of these factors remain unclear⁴⁹.

Correlation in nucleotide divergence. As seen in other mammalian genomes²³⁻²⁵, the average nucleotide divergence rate across 1-Mb windows varies significantly across the dog genome (coefficient of variation 0.11, compared with 0.024 expected under a uniform distribution). This regional variation shows significant correlation in orthologous windows across the dog, human and mouse genomes, but the strength of the correlation seems to decrease with total branch length (pair-wise correlation for orthologous 1-Mb windows: Spearman's $\rho = 0.49$ for dog-human and 0.24 for dog-mouse comparisons). Lineage-specific variation in the regional divergence

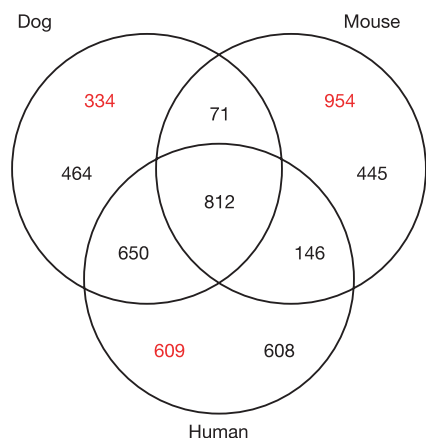


Figure 2 | Venn diagram showing the total lengths of aligned and unique sequences in the euchromatic portions of the dog, human and mouse genomes. Lengths shown in Mb, as inferred from genome-wide BLASTZ alignments (see Methods and Supplementary information). Overlapping partitions represent orthologous ancestral sequences. Each lineage-specific partition is further split into the total length of sequence classified as either lineage-specific interspersed repeats (red) or ancestral sequence (black). The latter is assumed to primarily represent ancestral sequences deleted in the two other species.

rates may be coupled with changes in factors such as sequence composition or chromosomal position^{23,54}. Consistent with this, the ratios of lineage-specific divergence rates in orthologous windows are positively correlated with the ratios of current G+C content in the same windows (Spearman's $\rho = 0.16$ for dog-human, 0.24 for dog-mouse).

Male mutation bias. Comparison of autosomal and X chromosome substitution rates can be used to estimate the relative mutation rates in the male and female germ lines (α), because the X chromosome is present in females twice as often as in males. Using the lineage-specific rates from ancestral repeats, we estimate α as 4.8 for the lineage leading to human, and 2.8 for the lineages leading to both mouse and dog. These values fall between recent estimates from murids^{24,25} and from hominids²³, and suggest that male mutation bias may have increased in the lineage leading to humans.

Mutational hotspots and chromosomal fission. Genome comparisons of human with both chicken⁵⁵ and chimpanzee²³ have previously revealed that sequences close to a telomere tend to have increased divergence rates and G + C content relative to interstitial sequences. It has been unclear whether these increases are inherent characteristics of the subtelomeric sequence itself or derived characteristics causally connected with its chromosomal position. We find a similar increase in both divergence (median increase 15%, $P < 10^{-5}$; Mann-Whitney U -test) and G+C content (median increase 9%, $P < 10^{-9}$) for subtelomeric regions along the dog lineage, with a sharp increase towards the telomeres (Supplementary Fig. S7).

This phenomenon is manifested at other synteny breaks, not only those at telomeres. We also observed a significant increase in divergence and G+C content in interstitial regions that are sites of syntenic breakpoints^{54,56} (Supplementary Fig. S7). These properties therefore seem correlated with the susceptibility of regions to chromosomal breakage.

Proportion of genome under purifying selection

One of the striking discoveries to emerge from the comparison of the human and mouse genomes^{21,24} was the inference that ~5.2% of the human genome shows greater-than-expected evolutionary conservation (compared with the background rate seen in ancestral repeat elements, which are presumed to be nonfunctional). This proportion greatly exceeds the 1–2% that can be explained by protein-coding regions alone. The extent and function of the large fraction of non-coding conserved sequence remain unclear⁵⁷, but this sequence is likely to include regulatory elements, structural elements and RNA genes.

Low turnover of conserved elements. We repeated the analysis of conserved elements using the human and dog genomes. Briefly, the

analysis involves calculating a conservation score S_{HD} , normalized by the regional divergence rate, for every 50-bp window in the human genome that can be aligned to dog. The distribution of conservation scores for all genomic sequences is compared to the distribution in ancestral repeat sequences (which are presumed to diverge at the local neutral rate), showing a clear excess of sequences with high conservation scores. By subtracting a scaled neutral distribution from the total distribution, one can estimate the distribution of conservation scores for sequences under purifying selection. Moreover, for a given sequence with conservation score S_{HD} , one can also assign a probability $P_{\text{selection}}(S_{HD})$ that the sequence is under purifying selection (see ref. 24 and Supplementary Information).

The human-dog genome comparison indicates that ~5.3% of the human genome is under purifying selection (Fig. 4a), which is equivalent to the proportion estimated from human-rodent analysis. The obvious question is whether the bases conserved between human and dog coincide with the bases conserved between humans and rodents^{25,58}. Because the conservation scores do not unambiguously assign sequences as either selected or neutral (but instead only assign probability scores for selection), we cannot directly compare the conserved bases. We therefore devised the following alternative approach.

We repeated the human-dog analysis, dividing the 1462 Mb of orthologous sequence between human and dog into those regions with (812 Mb) or without (650 Mb) orthologous sequence in mouse (Fig. 2). The first set shows a clear excess of conservation relative to background, corresponding to ~5.2% of the human genome (Fig. 4b). In contrast, the second set shows little or no excess conservation, corresponding to at most 0.1% of the human genome (Fig. 4c). This implies that hardly any of the functional elements conserved between human and dog have been deleted in the mouse lineage (see also Supplementary Information).

The results strongly suggest that there is a common set of functional elements across all three mammalian species, corresponding to ~5% of the human genome (~150 Mb). These functional elements reside largely within the 812 Mb of ancestral sequence common to human, mouse and dog. If we eliminate ancestral repeat elements within this shared sequence as largely non-functional, most functional elements can be localized to 634 Mb, and constitute approximately 24% of this sequence.

It should be noted that the estimate of ~5% pertains to conserved elements across distantly related mammals. It is possible that there are additional weakly constrained or recently evolved elements within narrow clades (for example, primates) that can only be detected by genomic sequencing of more closely related species²⁹.

Clustering of highly conserved non-coding elements. We next

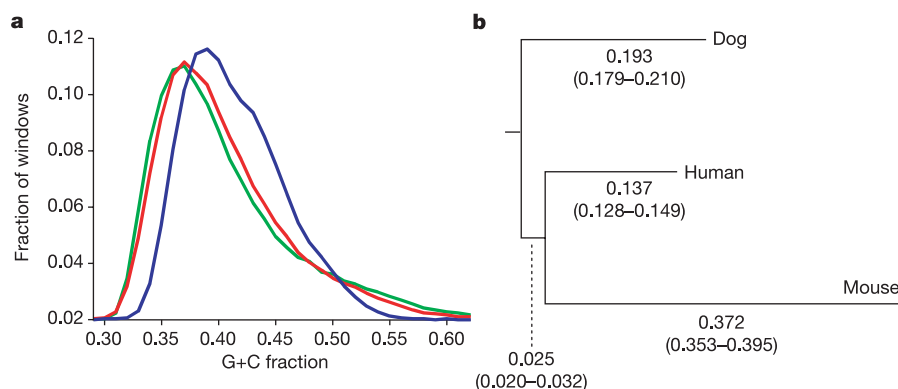


Figure 3 | Sequence composition and divergence rates. **a**, Distribution of G + C content in 10-kb windows across the genome in dog (green), human (red) and mouse (blue). **b**, Median lineage-specific substitution rates based on analysis of ancestral repeats aligning across all three genomes. Analysis was performed in non-overlapping 1-Mb windows across the dog genome

that contained at least 2 kb of aligned ancestral repeat sequence (median 8.8 kb). The tree was rooted with the consensus sequences from the ancestral repeats. Numbers in parentheses give the 20–80th percentile range across the windows studied.

explored the distribution of conserved non-coding elements (CNEs) across mammalian genomes. For this purpose, we calculated a conservation score S_{HMD} based on simultaneous conservation across all three species (see Methods). We defined highly conserved non-coding elements (HCNEs) to be 50-bp windows that do not overlap coding regions and for which $P_{\text{selection}}(S_{\text{HMD}})$, the probability of being under purifying selection given the conservation score, is at least 95%. We identified $\sim 140,000$ such windows (6.5 Mb total sequence), comprising $\sim 0.2\%$ of the human genome and representing the most conserved $\sim 5\%$ of all mammalian CNEs.

The density of HCNEs shows striking peaks when plotted in 1-Mb windows across the genome (Fig. 4d and Supplementary Figs S8 and S9), with 50% lying in 204 regions that span less than 14% of the human genome (Supplementary Table S11). These regions are generally gene-poor, together containing only $\sim 6\%$ of all protein-coding sequence.

The genes contained within these gene-poor regions are of particular interest. At least 182 of the 204 regions contain genes with key roles in establishing or maintaining cellular 'state'. At least 156 of the regions contain one or, in a few cases, several transcription factors involved in differentiation and development⁵⁹. Another 26 regions contain a gene important for neuronal specialization and growth, including several axon guidance receptors. The proportion of developmental regulators is far greater than expected by chance ($P < 10^{-31}$; see Supplementary Information).

We then tested whether the HCNEs within these regions tend to cluster around the genes encoding regulators of development. Analysis of the density of HCNEs in the intronic and intergenic sequences flanking every gene in the 204 regions revealed that the 197 genes encoding developmental regulators show an average of ~ 10 -fold enrichment for HCNEs relative to the full set of 1,285 genes

in the regions (Fig. 4e and Supplementary Fig. S10). The enrichment sometimes extends into the immediately flanking genes.

We note that the 204 regions include nearly all of the recently identified clusters of conserved elements between distantly related vertebrates such as chicken and pufferfish^{55,59–62}. For example, they overlap 56 of the 57 large intervals containing conserved non-coding sequence identified between human and chicken⁵⁵. The mammalian analysis, however, detects vastly more CNEs (>100 -fold more sequence than with pufferfish⁵⁹ and 2–3-fold more than with chicken) and identifies many more clusters. The limited sensitivity of these more distant vertebrate comparisons may reflect the difficulty of aligning short orthologous elements across such large evolutionary distances or the emergence of mammal-specific regulatory elements. In any case, mammalian comparative analysis may be a more powerful tool for elucidating the regulatory controls across these important regions.

Although the function of conserved non-coding elements is unknown, on the basis of recent studies^{59,63–66} it seems likely that many regulate gene expression. If so, the above results suggest that $\sim 50\%$ of all mammalian HCNEs may be devoted to regulating $\sim 1\%$ of all genes. In fact, the distribution may be even more skewed, as there are additional genomic regions with only slightly lower HCNE density than the 204 studied above (Supplementary Fig. S8). All of these regions clearly merit intensive investigation to assess indicators of regulatory function. We speculate that these regions may harbour characteristic chromatin structure and modifications that are potentially involved in the establishment or maintenance of cellular state.

Genes

Accurate identification of the protein-coding genes in mammalian genomes is essential for understanding the human genome, including its cellular components, regulatory controls and evolutionary

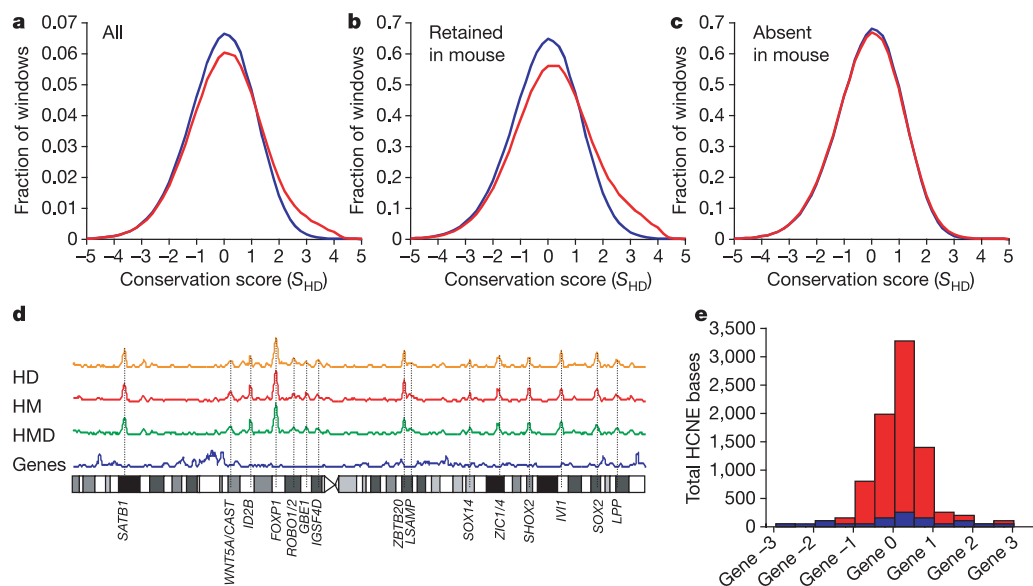


Figure 4 | Conservation of orthologous sequence between human and dog. **a**, Histogram of conservation scores, S , for all 50-bp windows across the human genome with at least 20 bases of orthologous sequence aligning to the dog genome, for all aligning sequences (red) and for ancestral repeat sequence only (blue). **b**, Conservation scores for the subset of windows that also have at least 20 bases of orthologous sequence aligning to the mouse genome. **c**, Conservation scores of the complementary subset of windows lacking such orthologous sequence in mouse. **d**, Density of 50-bp windows not overlapping known coding regions, for which $P_{\text{selection}}(S) > 95\%$, based on comparisons between human and dog (HD), human and mouse (HM), or between human, mouse and dog (HMD), and the density of known genes, all in 1-Mb sliding windows across human chromosome 3. **e**, Enrichment of

HCNEs in the immediate neighbourhood of genes encoding developmental regulators in the 204 highly conserved regions. The histogram shows the median number of HCNE bases in the intronic and surrounding intergenic sequence, for the 197 known or putative developmental regulators (indicated by top of red bar) and for all of the 1,285 genes (blue bar). The histogram is centred at the 5'-end of the gene (marked 0) and each bin corresponds to half of the normalized distance to the flanking consecutive upstream genes (marked -1, -2 and -3) or consecutive downstream genes (1, 2 and 3) as indicated. The sequences surrounding the developmental genes are typically longer, have more HCNE sequence and have a higher density of HCNE sequence than other genes in the regions (see Supplementary Information).

constraints. The number of protein-coding genes in human has been a topic of considerable debate, with estimates steadily falling from ~100,000 to 20,000–25,000 over the past decade^{21,22,67–70}. We analysed the dog genome in order to refine the human gene catalogue and to assess the evolutionary forces shaping mammals. (In the Genes section, 'gene' refers only to a protein-coding gene.)

Gene predictions in dog and human. We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains 19,300 dog gene predictions, with nearly all being clear homologues of known human genes.

The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (Ensembl build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable to spurious gene predictions in the human genome (M. Clamp, personal communication).

Gene duplications. Gene duplication is thought to contribute substantially to functional innovation^{69,71}. We identified 216 gene duplications that are specific to the dog lineage and 574 that are specific to the human lineage, using the synonymous substitution rate K_S as a distance metric and taking care to discard likely pseudogenes. (The CanFam 2.0 assembly contains approximately 24 additional gene duplications, mostly olfactory receptors.) Human genes are thus 2.7-fold more likely to have undergone duplication than are dog genes over the same time period. This may reflect increased repeat-mediated segmental duplication in the human lineage⁷².

Although gene duplication has been less frequent in dog than human, the affected gene classes are very similar. Prominent among the lineage-specific duplicated genes are genes that function in adaptive immunity, innate immunity, chemosensation and reproduction, as has been seen for other mammalian genomes^{24,25,69,71}. Reproductive competition within the species and competition against parasites have thus been major driving forces in gene family expansion.

The two gene families with the largest numbers of dog-specific genes are the histone H2B family and the α -interferons, which cluster in monophyletic clades when compared to their human homologues. This is particularly notable for the α -interferons, for which the gene families within the six species (human, mouse, rat, dog, cat and horse) are apparently monophyletic. This may be due either to coincidental independent gene duplication in each of the six lineages or to ongoing gene conversion events that have homogenized ancestral gene duplicates⁷³.

Evolution of orthologous genes across three species. The dog genome sequence allows us for the first time to characterize the large-scale patterns of evolution in protein-coding genes across three major mammalian orders. We focused on a subset of 13,816 human, mouse and dog genes with 1:1:1 orthology. For each, we inferred the number of lineage-specific synonymous (K_S) and non-synonymous (K_A) substitutions along each lineage and calculated the K_A/K_S ratio (Table 2 and Supplementary Information), a traditional measure of the strength of selection (both purifying and directional) on proteins⁷⁴.

The median K_A/K_S ratio differs sharply across the three lineages ($P < 10^{-44}$, Mann-Whitney U -test), with the dog lineage falling

between mouse and human. Population genetic theory predicts⁷⁵ that the strength of purifying selection should increase with effective population size (N_e). The observed relationship (mouse < dog < human) is thus consistent with the evolutionary prediction, given the expectation that smaller mammals tend to have larger effective population sizes⁷⁶.

We next searched for particular classes of genes showing deviations from the expected rate of evolution for a species. Such variation in rate (heterotachy) may point to lineage-specific positive selection or relaxation of evolutionary constraints⁷⁷. We developed a statistical method similar to the recently described Gene Set Enrichment Analysis (GSEA)^{78–80} to detect evidence of heterotachy for sets of functionally related genes (see Supplementary Information). Briefly, the approach involves ranking all genes by K_A/K_S ratio, testing whether the set is randomly distributed along the list and assessing the significance of the observed deviations by comparison with randomly permuted gene sets. In contrast to previous studies, which focused on small numbers of genes with prior hypotheses of selection, this approach detects signals of lineage-specific evolution in a relatively unbiased manner and can provide context to the results of more limited studies.

A total of 4,950 overlapping gene sets were studied, defined by such criteria as biological function, cellular location or co-expression (see Supplementary Information). Overall, the deviations between the three lineages are small, and median K_A/K_S ratios for particular gene sets are highly correlated for each pair of species (Supplementary Fig. S11). However, there is greater relative variation in human–mouse and dog–mouse comparisons than in human–dog comparisons (Supplementary Fig. S12).

This suggests that observed heterotachy between human and mouse must be interpreted with caution. For example, there is a great interest in the identification of genetic changes underlying the unique evolution of the human brain. A recent study⁸¹ highlighted 24 genes involved in brain development and physiology that show signs of accelerated evolution in the lineage leading from ancestral primates to humans when compared to their rodent orthologues. We observe the same trend for the 18 human genes that overlap with the genes studied here, but find at least as many genes with higher relative acceleration in the dog lineage (see Supplementary Information). Heterotachy relative to mouse therefore does not appear to be a distinctive feature of the human lineage. It may reflect decelerated evolution in the rodent lineage, or possibly independent adaptive evolution in the human and dog lineages⁸².

A small number of gene sets show evidence of significantly accelerated evolution in the human lineage, relative to both mouse and dog (32 sets at $z \geq 5.0$ versus zero sets expected by chance, $P < 10^{-4}$; Fig. 5a). These sets fall into two categories: genes expressed exclusively in testis, and (nuclear) genes encoding subunits of the mitochondrial electron transport chain (ETC) complexes. The former are believed to undergo rapid evolution as a consequence of sperm competition across a wide range of species^{83–85}, and lineage-specific acceleration suggests that sexual selection may have been a particularly strong force in primate evolution. The selective forces acting on the latter category are less obvious. Because of the importance of mitochondrial ATP generation for sperm motility⁸⁶, and the potentially antagonistic co-evolution of these genes with maternally inherited mitochondrial DNA-encoded subunits⁸⁷, we

Table 2 | Evolutionary rates for 1:1:1 orthologues among dog, mouse and human

	Median (20–80th percentile range)			Spearman's ρ		
	Dog*	Mouse	Human	Dog-human	Dog-mouse	Human-mouse
K_S	0.210 (0.138–0.322)	0.416 (0.310–0.558)	0.139 (0.0928–0.214)	0.47	0.50	0.52
K_A	0.021 (0.006–0.051)	0.038 (0.013–0.087)	0.017 (0.005–0.040)	0.87	0.87	0.86
K_A/K_S	0.095 (0.030–0.221)	0.088 (0.031–0.197)	0.112 (0.034–0.272)	0.80	0.85	0.82

*Estimates are based on unrooted tree. The dog branch thus includes the branch from the boreoeutherian ancestor to the primate–rodent split.

propose that sexual selection may also be the primary force behind the rapid evolution of the primate ETC genes. Given the ubiquitous role of mitochondrial function, however, such sexual selection may have led to profound secondary effects on physiology⁸⁸.

We found no gene sets with comparably strong evidence for dog-specific accelerated evolution. There is, however, a small excess of sets with moderately high acceleration scores (19 sets at $z \geq 3.0$ versus 5 sets expected by chance, $P < 0.02$; Fig. 5b). These sets, which are primarily related to metabolism, may contain promising candidates for follow-up studies of molecular adaptation in carnivores.

Polymorphism and haplotype structure in the domestic dog

The modern dog has a distinct population structure with hundreds of genetically isolated breeds, widely varying disease incidence and distinctive morphological and behavioural traits^{89,90}. Unlocking the full potential of the dog genome for genetic analysis requires a dense SNP map and an understanding of the structure of genetic variation both within and among breeds.

Generating a SNP map. We generated a SNP map of the dog genome containing >2.5 million distinct SNPs mapped to the draft genome sequence, corresponding to an average density of approximately one SNP per kb (Table 3). The SNPs were discovered in three complementary ways (see Supplementary Information). (1) We identified SNPs within the sequenced boxer genome (set 1; $\sim 770,000$ SNPs) by searching for sites at which alternative alleles are supported by at least two independent reads each. We tested a subset ($n = 40$ SNPs) by genotyping and confirmed all as heterozygous sites. (2) We compared the 1.5 $\times$ sequence from the standard poodle¹⁶ with the draft genome sequence from the boxer (set 2; $\sim 1,460,000$ SNPs). (3) We generated shotgun sequence data from nine diverse dog breeds ($\sim 100,000$ reads each, 0.02 $\times$ coverage), four grey wolves and one coyote ($\sim 22,000$ reads each, 0.004 $\times$ coverage) and compared it to the boxer (set 3; $\sim 440,000$ SNPs). We tested a subset ($n = 1,283$ SNPs) by genotyping and confirmed 96% as true polymorphisms.

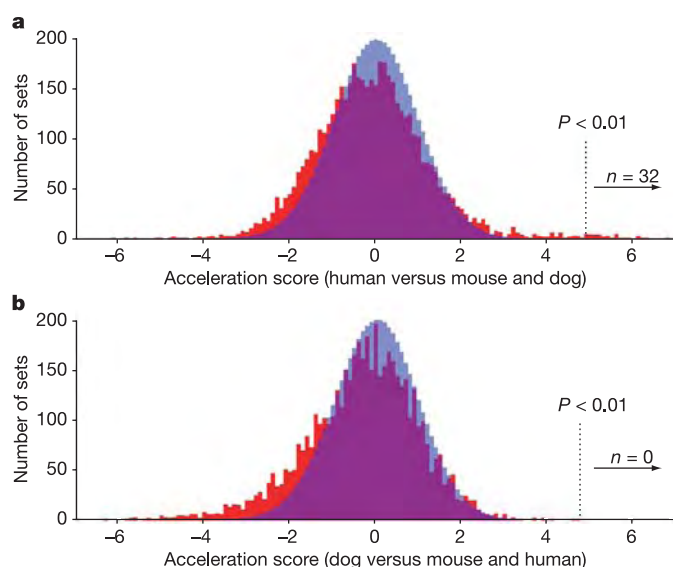


Figure 5 | Gene sets showing accelerated evolution along the human and dog lineages. **a**, Distribution of acceleration scores along the human lineage relative to both mouse and dog, observed for 4,950 gene sets (red). The expected distribution based on 10,000 randomized trials is shown in blue. The dotted line shows the acceleration score for which the probability of observing even a single set by random chance (out of the 4,950 sets tested) is less than 1%. In fact, 32 sets show acceleration scores on the human lineage exceeding this threshold. **b**, The observed (red) and expected (blue) distribution of acceleration scores for the dog lineage, relative to both human and mouse.

Table 3 | SNPs discovered in dogs, wolves and coyotes compared to the boxer assembly

Set number	Breed or species	Number of SNPs	SNP rate (one per x bases)
1	Boxer versus boxer	768,948	3,004 (observed) 1,637 (corrected)
2	Boxer versus poodle	1,455,007	894
3a	Boxer versus breeds*		
	German shepherd	45,271	900
	Rottweiler	44,097	917
	Bedlington terrier	44,168	913
	Beagle	42,572	903
	Labrador retriever	40,730	926
	English shepherd	40,935	907
	Italian greyhound	39,390	954
	Alaskan malamute	45,103	787
	Portuguese water dog	45,457	896
	Total distinct SNPs	373,382	900
3b	Boxer versus Canids†		
	China grey wolf	12,182	580
	Alaska grey wolf	13,888	572
	India grey wolf	14,510	573
	Spanish grey wolf	10,349	587
	California coyote	20,270	417
	Total distinct SNPs	71,381	
3	Set 3 total distinct SNPs	441,441	
Total	Total distinct SNPs	2,559,519	

* Based on $\sim 100,000$ sequence reads per breed.

† Based on $\sim 20,000$ sequence reads per wolf.

The SNP rate between the boxer and any of the different breeds is one SNP per ~ 900 bp, with little variation among breeds (Table 3). The only outlier ($\sim 1/790$ bp) is the Alaskan malamute, which is the only breed studied that belongs to the Asian breed cluster⁹¹. The grey wolf ($\sim 1/580$ bp) and coyote ($\sim 1/420$ bp) show greater variation when compared with the boxer, supporting previous evidence of a bottleneck during dog domestication, whereas that the SNP rate is lower in the grey wolf than in the coyote reflects the closer relationship of the grey wolf to the domestic dog^{1-3,92} (see section 'Resolving canid phylogeny').

The observed SNP rate within the sequenced boxer assembly is $\sim 1/3,000$ bp. This underestimates the true heterozygosity owing to the conservative criterion used for identifying SNPs within the boxer assembly (requiring two reads containing each allele); correcting for this leads to an estimate of $\sim 1/1,600$ bp (see Supplementary Information). This low rate reflects reduced polymorphism within a breed, compared with the greater variation of $\sim 1/900$ bp between breeds.

To assess the utility of the SNPs for dog genetics, we genotyped a subset from set 3a ($n = 1,283$) in 20 dogs from each of ten breeds (Supplementary Table S16). Within a typical breed, $\sim 73\%$ of the SNPs were polymorphic. The polymorphic SNPs have minor allele frequencies that are approximately evenly distributed between 5% and 50% (allele frequencies less than 5% are not reliable with only 40 chromosomes sampled). In addition, the SNPs from sets 2 and 3 have a roughly uniform distribution across the genome (Fig. 6a, see below concerning set 1). The SNP map thus has high density, even distribution and high cross-breed polymorphism, indicating that it should be valuable for genetic studies.

Expectations for linkage disequilibrium and haplotype structure.

Modern dog breeds are the product of at least two population bottlenecks, the first associated with domestication from wolves ($\sim 7,000$ – $50,000$ generations ago) and the second resulting from intensive selection to create the breed (~ 50 – 100 generations ago). This population history should leave distinctive signatures on the patterns of genetic variation both within and across breeds. We might expect aspects of both the long-range LD seen in inbred mouse strains, with strain-specific haplotypes extending over multiple megabases, and the short-range LD seen in humans, with ancestral haplotype blocks typically extending over tens of kilobases. Specifically,

long-range LD would be expected within dog breeds and short-range LD across breeds.

Preliminary evidence of long-range LD within breeds has been reported⁹⁰. Five genome regions were examined (~1% of the genome) in five breeds using ~200 SNPs with high minor allele frequency. LD seemed to extend 10–100-fold further in dog than in human, with relatively few haplotypes per breed.

With the availability of a genome sequence and a SNP map, we sought to undertake a systematic analysis of LD and haplotype structure in the dog genome.

Haplotype structure within the boxer assembly. We first analysed the structure of genetic variation within the sequenced boxer genome by examining the distribution of the ~770,000 SNPs detected between homologous chromosomes. Strikingly, the genome is a mosaic of long, alternating regions of near-total homozygosity and high heterozygosity (Fig. 6b, c), with observed SNP rates of ~14 per Mb and ~850 per Mb, respectively. (The latter is close to that seen within breeds and is indistinguishable when one corrects for the conservative criterion used to identify SNPs within the boxer assembly; see Supplementary Information.) The homozygous regions have an N50 size of 6.9 Mb and cover 62% of the genome, and the heterozygous regions have an N50 size of 1.1 Mb and cover

38% of the genome. The results imply that the boxer genome is largely comprised of vast haplotype blocks. The long stretches of homozygosity indicate regions in which the sequenced boxer genome carries the same haplotype on both chromosomes. The proportion of homozygosity (~62%) reflects the limited haplotype diversity within breeds.

Long-range haplotypes in different breeds. We sought to determine whether the striking haplotype structure seen in the boxer genome is representative of most dog breeds. To this end, we randomly selected ten regions of 15 Mb each (~6% of the genome) and examined linkage disequilibrium in these regions in a collection of 224 dogs, consisting of 20 dogs from each of ten breeds and one dog from each of 24 additional breeds (see Supplementary Tables S17–S19).

The ten breeds were chosen to represent all four clusters described in ref. 91. The selected breeds have diverse histories, with varying population size and bottleneck severity. For example, the Basenji is an ancient breed from Africa that has a small breeding population in the United States descending from dogs imported in the 1930s–1940s (refs 93, 94). The Irish wolfhound suffered a severe bottleneck two centuries ago, with most dogs today being descendants of a single dog in the early 1800s (refs 5, 94). In contrast, the Labrador retriever and golden retriever have long been, and remain, extremely popular dogs

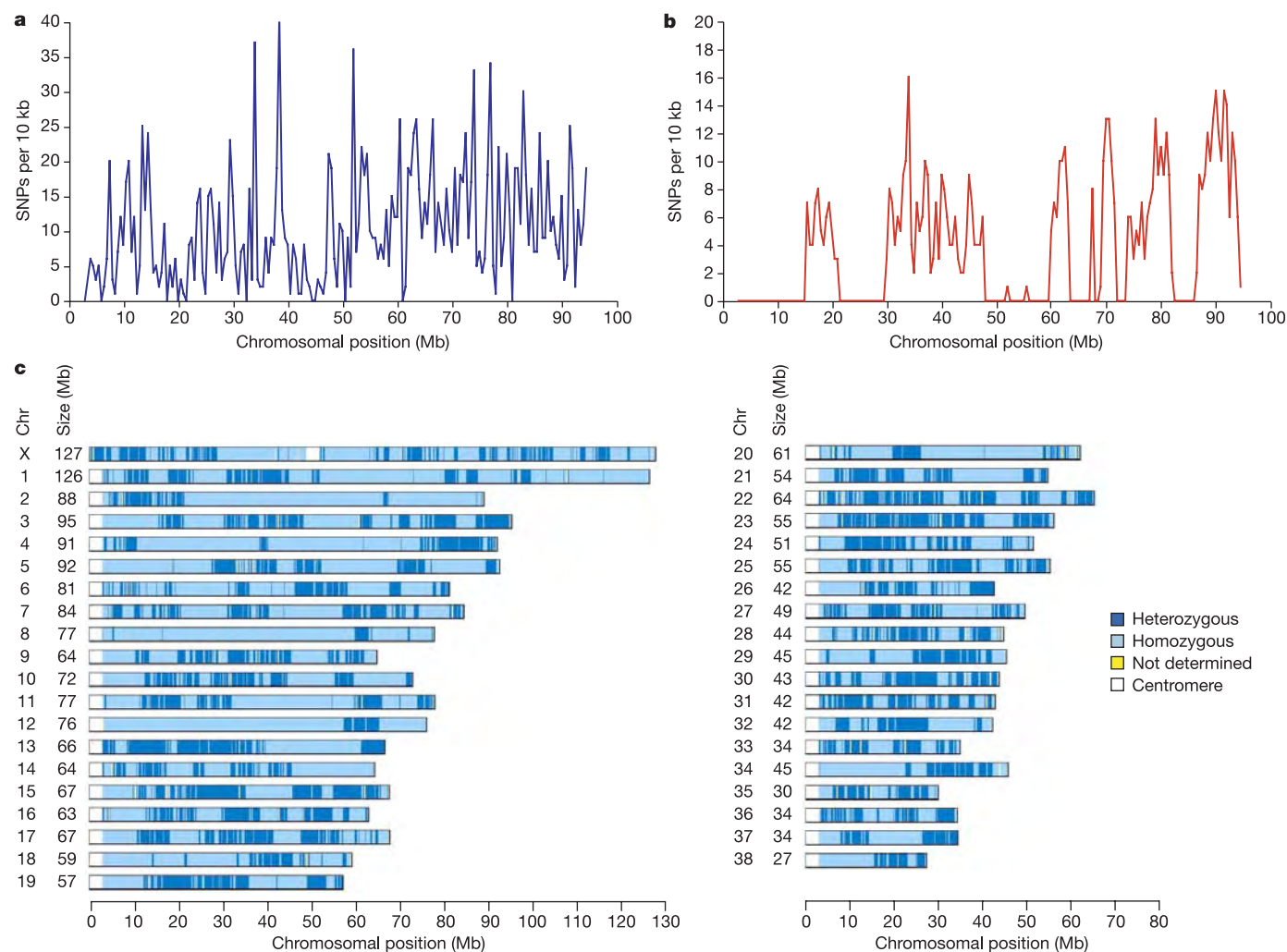


Figure 6 | The distribution of SNPs is fairly uniform across breeds, but non-uniform within the sequenced boxer assembly. **a**, SNPs across chromosome 3, generated by comparing the boxer assembly with WGS reads from nine breeds. **b**, The SNPs on chromosome 3 of the boxer assembly show an uneven distribution (plotted in 500-kb windows). Note that boxer SNPs were identified using a more conservative method, lowering the observed

SNP rate by roughly twofold. **c**, An alternating pattern of large homozygous (light blue, ~62% of genome; N50 size 6.9 Mb) and large heterozygous (dark blue ~38% of genome; N50 size 1.1 Mb) blocks indicates large identical or divergent haplotypes across the boxer genome. White indicates centromeric sequence.

(with ~150,000 and ~50,000 new puppies registered annually, respectively). They have not undergone such recent severe bottlenecks, but some lines have lost diversity because of the repeated use of popular sires⁸⁹. The Glen of Imaal terrier represents the opposite end of the popularity spectrum, with fewer than 100 new puppies registered with the American kennel Club each year.

The 224 dogs were genotyped for SNPs across each of the ten regions, providing 2,240 cases in which to assess long-range LD. The SNPs ($n = 1,219$; Supplementary Table S19) were distributed along the regions to measure the fall-off of genetic correlation, with higher density at the start of the region and lower densities at further distances (Fig. 7a). In 645 cases, we also examined the first 10 kb in

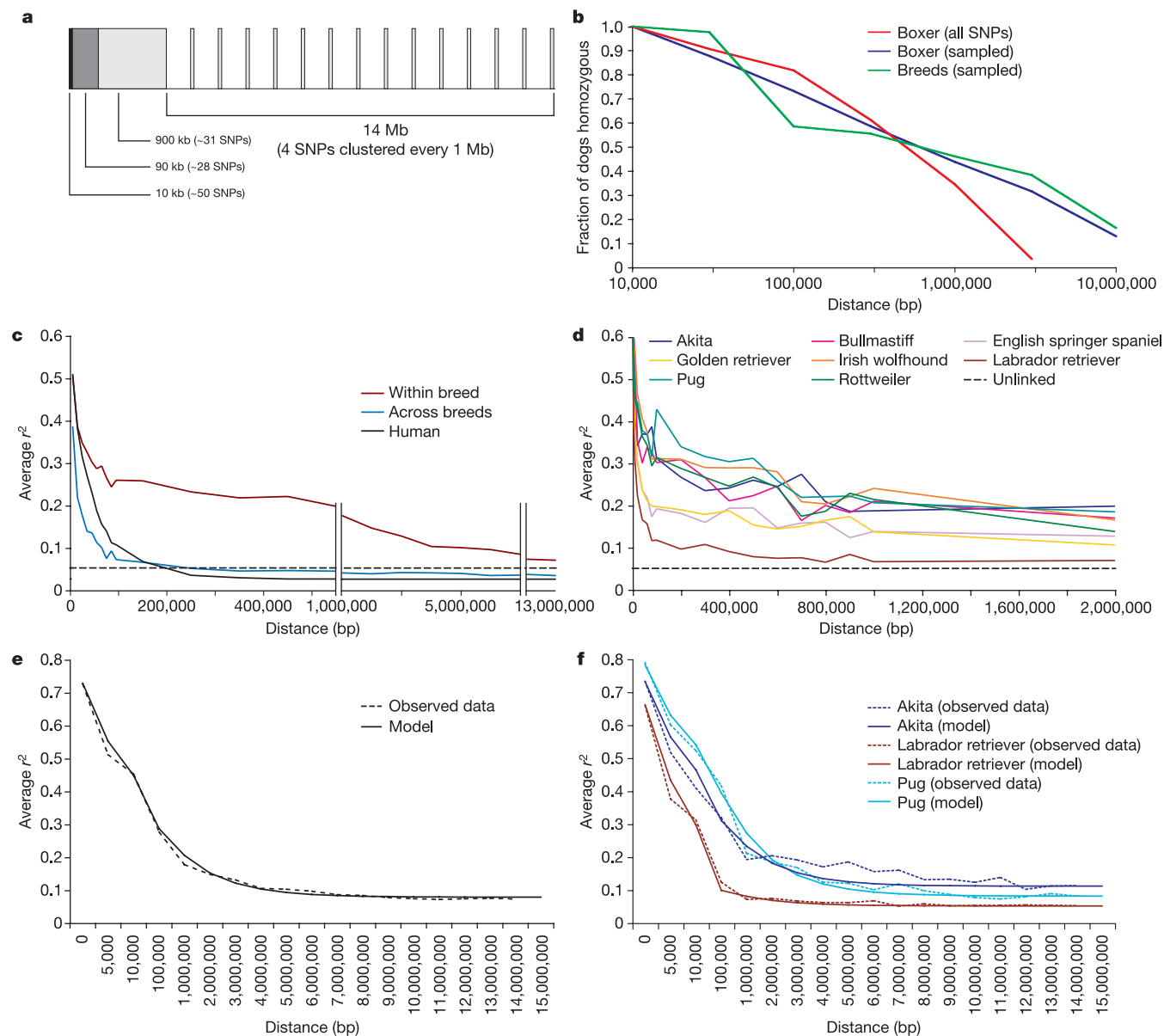


Figure 7 | Homozygous regions and linkage disequilibrium are nearly 100-fold longer within dog breeds than across the dog or human populations. **a**, Sampling design for ten random regions of 15 Mb each, used to assess the haplotype structure of ~6% of the genome (see Supplementary Information). For each region, we examined the first 10 kb through resequencing and dense genotyping. To detect long haplotypes, we genotyped SNPs distributed throughout the next 1 Mb and sampled SNPs at intervals of 1 Mb for the next 14 Mb. In total we genotyped 1,219 SNPs across the ten regions in a collection of 224 dogs (20 dogs from each of 10 breeds and one dog from each of 24 breeds). **b**, Conditional on a dog being homozygous for the initial 10-kb region ($n = 245$), we assessed the probability that the dog was homozygous for all SNPs within a given distance. The average proportion remaining homozygous is compared for the various breeds (green), for the boxer when sampled in the same ways as the breeds (blue) and for the boxer using all SNPs found in the genome sequence (red). About 50% of the individuals seem to be homozygous throughout 1 Mb both in the boxer and other breeds, indicating that other

breeds have comparable long-range homozygosity. **c**, Linkage disequilibrium (LD) as a function of distance is shown as the r^2 statistic within individual breeds (red), across various breeds (blue) and a human population (black) taken from the CEPH collection genotyped as part of the ENCODE component of the International HapMap Project¹⁸. For the overall dog and human populations, LD falls rapidly, reaching the baseline level seen for unlinked loci by ~200 kb. In contrast, LD for individual breeds falls initially but then stays at a moderately high level across several megabases. **d**, The LD curves are broadly similar for most breeds, but the proportion of long-range LD is correlated with known breed history. **e**, The observed within-breed LD curve (averaged across breeds) is well fitted by a simple model with a domestication bottleneck 10,500 generations ago and a breed-creation bottleneck occurring 50 generations ago (see Supplementary Information). **f**, LD curves for individual dog breeds can be fitted by models with different breed-creation bottlenecks. The poorest fit is obtained for the akita, the breeding history of which is known to involve two separate breed-creation bottlenecks.

greater detail by denser genotyping (with ~ 2 SNPs per kb) in 405 cases and complete resequencing in 240 cases. The resequencing data yielded a heterozygosity rate of ~ 1 SNP per 1,500 bp, essentially equivalent to the rate seen in the sequenced boxer genome.

On the basis of examining the first 10 kb, we found that $\sim 38\%$ of instances seem to be completely homozygous and that all dogs seem to be homozygous for at least one of the ten regions. We then measured the distance over which homozygosity persisted. Of instances homozygous in the initial 10-kb segment, 46% were homozygous across 1 Mb and 17% were still homozygous across 10 Mb (Fig. 7b). The fall-off in homozygosity is essentially identical to that seen in the boxer genome, provided that the boxer data are sampled in an equivalent manner (see Supplementary Information). This indicates that the long-range haplotype structure seen in the boxer is typical of most dog breeds, although the precise haplotypes vary with breed and the locations of homozygous regions vary between individuals.

We also assessed long-range correlations by calculating r^2 , a traditional measure of LD, across the 15-Mb regions. The r^2 curve representing the overall dog population (one dog from each of 24 breeds) drops rapidly to background levels. This is in sharp contrast to the r^2 curves within each breed. Within breeds, LD is biphasic, showing a sharp initial drop within ~ 90 kb followed by an extended shoulder that gradually declines to the background (unlinked) level by 5–15 Mb in most breeds (Fig. 7c). The basic pattern is similar in all ten regions (Supplementary Fig. S13) and in all breeds (Fig. 7d). (Labrador retrievers show the shortest LD, probably due to their mixed aetiology and large population size.)

The biphasic r^2 curves within each breed thus consist of two components (Fig. 7e), at scales differing by ~ 100 -fold. The first component matches the fall-off in the general dog population and is likely to represent the short-range de-correlation of local haplotype blocks in the ancestral dog population. The second component represents long-range breed-specific haplotypes (Fig. 8a). Notably, the first component falls off nearly twice as quickly as the LD in the human population (~ 200 kb), and the second component falls off slightly slower than seen in laboratory mouse strains⁹⁵.

Modelling the effects of population history. We tested this interpretation by performing mathematical simulations on a dog population that underwent an ancient bottleneck and recent breed-creation bottlenecks, using the coalescent approach⁹⁶ (see Supplementary Information). Our experimental results were well fitted by models assuming an ancient bottleneck (effective domesticated population size 13,000, inbreeding coefficient $F = 0.12$) occurring $\sim 9,000$ generations ago (corresponding to $\sim 27,000$ years) and subsequent breed-creation bottlenecks of varying intensities occurring 30–90 generations ago⁹⁷ (Supplementary Fig. S14). The model closely reproduces the observed r^2 curves and the observed polymorphism rates within breeds, among breeds and between dog and grey wolf. The model also yields estimates of breed-specific bottlenecks that are broadly consistent with known breed histories. For example, Labrador retrievers, and to a lesser extent golden retrievers and English springer spaniels, show less severe bottlenecks.

Deterministically modelled results (Fig. 7e, f) indicate that a simple, two-bottleneck model provides a close fit to the data for the breeds. They do not rule out a more complex population history, such as multiple domestication events, low levels of continuing gene flow between domestic dog and grey wolf^{97,98} or multiple bottlenecks within breeds. Notably, the akita yields the poorest fit to the model, with an r^2 curve that appears to be triphasic. This may reflect the initial creation of the breed as a hunting dog in Japan ~ 450 generations ago, and a consecutive bottleneck associated with its introduction into the United States during the 1940s (ref. 99).

Haplotype diversity. We next studied haplotype diversity within and among breeds, using the dense genotypes from the 10-kb regions. Across the 645 cases examined, there is an average of ~ 10 distinct haplotypes per region. Within a breed, we typically see four of

these haplotypes, with the average frequency of the most common haplotype being 55% and the average frequency of the two most common being 80% (Fig. 8c and Supplementary Fig. S18). The haplotypes and their frequencies differ sharply across breeds. Nonetheless, 80% of the haplotypes seen with a frequency of at least 5% in one breed are found in other breeds as well (Supplementary Table S26). This extends previous observations of haplotype sharing across breeds⁹⁰. In particular, the inclusion of all SNPs with a minor allele frequency $\geq 5\%$ across all breeds provides a more accurate picture of haplotype sharing, because the analysis includes haplotypes that are rare within a single breed but more common across the population.

We then inferred the ancestral haplotype block structure in the ancestral dog population (before the creation of modern breeds) by combining the data across breeds and applying methods similar to those used for haplotype analysis in the human genome¹⁰⁰ (see Supplementary Information). In the 10-kb regions studied, one or two haplotype blocks were typically observed. Additional data across 100-kb regions suggest that the ancestral blocks have an average size of ~ 10 kb. The blocks typically have ~ 4 –5 distinct haplotypes across the entire dog population (Fig. 8b). The overall situation closely resembles the structure for the human genome, although with slightly smaller block size (Supplementary Figs S15–S19 and Supplementary Table S24–S26).

Ancestral and breed-specific haplotypes. A clear picture of the population genetic history of dogs emerges from the results detailed above:

- The ancestral dog population had short-range LD. The haplotype blocks were somewhat shorter than in modern humans (~ 10 kb versus ~ 20 kb in human), consistent with the dog population being somewhat older than the human population ($\sim 9,000$ generations versus $\sim 4,000$ generations). Haplotype blocks at large distances were essentially uncorrelated (Fig. 8a).
- Breed creation introduced tight breed-specific bottlenecks, at least for the breeds examined. From the great diversity of long-range haplotype combinations carried in the ancestral population, the founding chromosomes emerging from the bottleneck represented only a small subset. These became long-range breed-specific haplotypes (Fig. 8a).
- Although the breed-specific bottlenecks were tight, they did not cause massive random fixation of individual haplotypes. Only 13% of the small ancestral haplotypes are monomorphic within a typical breed, consistent with the estimated inbreeding coefficient of $\sim 12\%$. Across larger regions (≥ 100 kb), we observed no cases of complete fixation within a breed (Supplementary Fig. S20).
- There is notable sharing of 100-kb haplotypes across breeds, with $\sim 60\%$ seen in multiple breeds although with different frequencies. On average, the probability of sampling the same haplotype on two chromosomes chosen from different breeds is roughly twofold lower than for chromosomes chosen within a single breed (Supplementary Fig. S21).

Implications for genetic mapping. These results have important implications for the design of dog genetic studies. Although early efforts focused on cross-breeding of dogs for linkage analysis^{101–103}, it is now clear that within-breed association studies offer specific advantages in the study of both monogenic and polygenic diseases. First, they use existing dogs coming to medical attention and do not require the sampling of families with large numbers of affected individuals. Such studies should be highly informative, because dog breeds have retained substantial genetic diversity. Moreover, they will require a much lower density of SNPs than comparable human association studies, because the long-range LD within breeds extends ~ 50 -fold further than in humans^{90,104,105}.

Whereas human association studies require $>300,000$ evenly spaced SNPs^{100,106,107}, the fact that LD extends over at least 50-fold greater distances in dog suggests that dog association studies would require perhaps $\sim 10,000$ evenly spaced SNPs. To estimate the

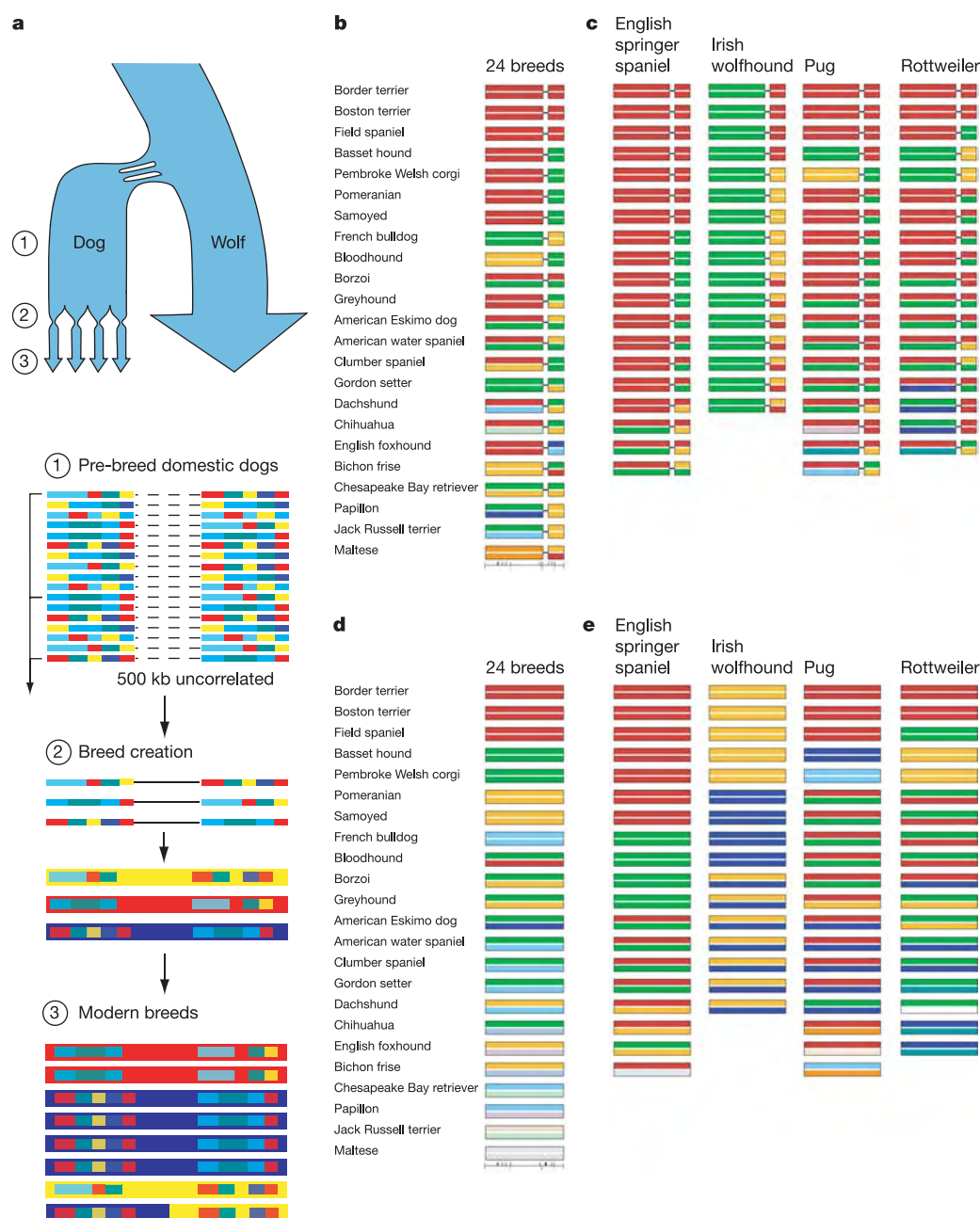


Figure 8 | Two bottlenecks, one old and one recent, have shaped the haplotype structure and linkage disequilibrium of canine breeds.

a, Modern haplotype structure arose from key events in dog breeding history. The domestic dog diverged from wolves 15,000–100,000 years ago^{97,119}, probably through multiple domestication events⁹⁸. Recent dog breeds have been created within the past few hundred years. Both bottlenecks have influenced the haplotype pattern and LD of current breeds. (1) Before the creation of modern breeds, the dog population had the short-range LD expected on the basis of its large size and time since the domestication bottleneck. (2) In the creation of modern breeds, a small subset of chromosomes was selected from the pool of domestic dogs. The long-range patterns that happened to be carried on these chromosomes became common within the breed, thereby creating long-range LD. (3) In the short time since breed creation, these long-range patterns have not yet been substantially broken down by recombination. Long-range haplotypes, however, still retain the underlying short-range ancestral haplotype blocks from the domestic dog population, and these are revealed when one examines chromosomes across many breeds. **b**, **c**, Distribution of ancestral haplotype blocks in a 10-kb window on chromosome 6 at ~31.4 Mb across

24 breeds (**b**) and within four breeds (**c**). Ancestral haplotype blocks are 5–15 kb in size (which is shorter than the ~25-kb blocks seen in humans) and are shared across breeds. Typical blocks show a spectrum of ~5 haplotypes, with one common major haplotype. Blocks were defined using the modified four-gamete rule (see Supplementary Information) and each haplotype (minor allele frequency (maf) > 3%) within a block was given a unique colour. **d**, **e**, Distribution of breed-derived haplotypes across a 10-kb window on chromosome 6 at ~31.4 Mb across 24 breeds (**d**) and within four breeds (**e**). Each colour denotes a distinct haplotype (maf > 3%) across 11 SNPs in the 10-kb window for each of the analysed dogs. Pairs of haplotypes have an average of 3.7 differences. Most haplotypes can be definitively identified on the basis of homozygosity within individual dogs. Grey denotes haplotypes that cannot be unambiguously phased owing to rare alleles or missing data. Within each of the four breeds shown, there are 2–5 haplotypes, with one or two major haplotypes accounting for the majority of the chromosomes. Across the 24 breeds, there are a total of seven haplotypes. All but three are seen in multiple breeds, although at varying frequencies.

number of SNPs required, we generated SNP sets from ten 1-Mb regions by coalescent simulations using the bottleneck parameters that generate SNP rates and LD curves equivalent to the actual data (Supplementary Fig. S14 and Supplementary Table S20). We then selected individual SNPs as 'disease alleles' and tested our ability to map them by association analysis with various marker densities (Fig. 9a).

For disease alleles causing a simple mendelian dominant trait with high penetrance and no phenocopies, there is overwhelming power to map the locus (Fig. 9a). Using ~15,000 evenly spaced SNPs and a log likelihood odds ratio (LOD score) score threshold of 5, the probability of detecting the locus is over 99% given a collection of 100 affected and 100 unaffected dogs. (The LOD score threshold corresponds to a false positive rate of 3% loci per genome.)

For a multigenic trait, the power to detect disease alleles depends on several factors, including the relative risk conferred by the allele, the allele frequency and the interaction with other alleles. We investigated a simple model of an allele that increases risk by a multiplicative factor (λ) of 2 or 5 (see Supplementary Information). Using the above SNP density and LOD score threshold, the power to detect a locus with a sample of 100 affected and 100 unaffected dogs is 97% for $\lambda = 5$ and 50% for $\lambda = 2$ (Fig. 9b, c). Although initial mapping will be best done by association within breeds, subsequent fine-structure mapping to pinpoint the disease gene will probably benefit from cross-breed comparison. Given the genetic relationships across breeds described above, it is likely that the same risk allele will be carried in multiple breeds. By comparing risk-associated haplotypes in multiple breeds, it should be possible to substantially narrow the region containing the gene.

Resolving canid phylogeny

The dog family, Canidae, contains 34 closely related species that diverged within the last ~10 million years¹. Resolving the evolutionary relationships of such closely related taxa has been difficult because a great quantity of genomic sequence is typically required to yield enough informative nucleotide sites for the unambiguous reconstruction of phylogenetic trees. We sought to streamline the process of evolutionary reconstruction by exploiting our knowledge of the dog genome to select genomic regions that would maximize the amount of phylogenetic signal per sequenced base. Specifically, we sought regions of rapidly evolving, unique sequence.

We first compared the coding regions of 13,816 dog genes with human–dog–mouse 1:1:1 orthologues to find those with high neutral evolutionary divergence (comparing K_S and K_A/K_S). We selected 12 exons (8,080 bp) for sequencing, based on the criteria that their sequences (1) are consistent with the known phylogeny of human,

dog, mouse and rat, (2) have a high percentage of bases ($\geq 15\%$) that are informative for phylogenetic reconstruction in the human, dog, mouse and rat phylogenies, and (3) could be successfully amplified in all canids. The chosen exons contain 3.3-fold more substitutions than random exonic sequence. Using our SNP database, we also evaluated introns to identify those with high variation between dog and coyote. We selected four introns (3,029 bp) that contained ~5-fold more SNPs than the background frequency. We sequenced these exons and introns (11,109 bp) in 30 out of 34 living wild canids, and we combined the data with additional sequences (3,839 bp) from recent studies^{3,92}.

The resulting evolutionary tree has a high degree of statistical support (Fig. 10), and uniquely resolves the topology of the dog's closest relatives. Grey wolf and dog are most closely related (0.04% and 0.21% sequence divergence in nuclear exon and intron sequences, respectively), followed by a close affiliation with coyote, golden jackal and Ethiopian wolf, three species that can hybridize with dogs in the wild (Fig. 10). Closest to this group are the dhole and African wild dog, two species with a uniquely structured meat-slicing tooth, suggesting that this adaptation was later lost. The molecular tree supports an African origin for the wolf-like canids, as the two African jackals are the most basal members of this clade. The two other large groupings of canids are (1) the South American canids, which are clearly rooted by the two most morphologically divergent canids, the maned wolf and bush dog; and (2) the red fox-like canids, which are rooted by the fennec fox and Blanford's fox, but now also include the raccoon dog and bat-eared fox with higher support. Together, these three clades contain 93% of all living canids. The grey fox lineage seems to be the most primitive and suggests a North American origin of the living canids about 10 million years ago¹.

These results demonstrate the close kinship of canids. Their limited sequence divergence suggests that many molecular tools developed for the dog (for example, expression microarrays) will be useful for exploring adaptation and evolutionary divergence in other canids as well.

Conclusions

Genome comparison is a powerful tool for discovery. It can reveal unknown—and even unsuspected—biological functions, by sifting the records of evolutionary experiments that have occurred over 100 years or over 100 million years. The dog genome sequence illustrates the range of information that can be gleaned from such studies.

Mammalian genome analysis is helping to develop a global picture of gene regulation in the human genome. Initial comparison with rodents revealed that ~5% of the human genome is under purifying selection, and that the majority of this sequence is not protein-

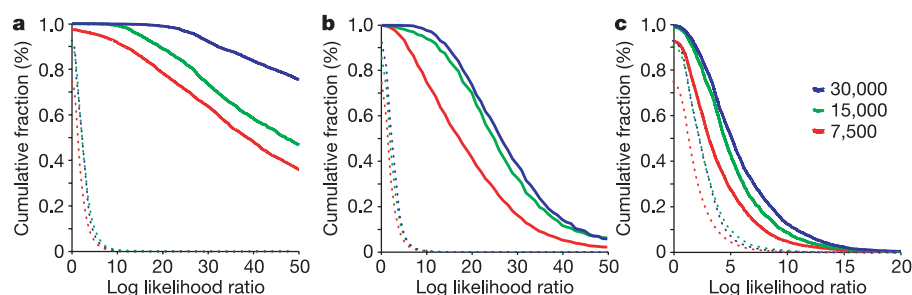


Figure 9 | Power to detect a disease locus by association mapping. One SNP was designated as a disease allele under one of three genetic models: (a) simple mendelian dominant, (b) fivefold multiplicative increase in risk and (c) twofold multiplicative increase in risk. SNP genotypes across surrounding chromosomal regions of 1 Mb were simulated, using the coalescent model corresponding to observed within-breed variation (see text). Diploid genotypes across the chromosomal region were then generated for 100 affected and 100 unaffected dogs, based on the disease model, and association analysis was performed to detect the presence of the

disease allele. The distribution of the maximum LOD score across the 1-Mb region is shown for analyses based on multi-SNP haplotypes (solid lines) with SNP densities equivalent to a genome-wide map with a total of 7,500 (red), 15,000 (green) or 30,000 (blue) SNPs. Dotted curves show the null distribution for a genome-wide search in which no disease locus is present (see Supplementary Information). A LOD score of 5 corresponds to <3% chance of a false positive across the genome. For this threshold, the power to detect a disease allele that increases risk by twofold using haplotype analysis and a map with 15,000 SNPs is ~50%.

coding. The dog genome is now further clarifying this picture, as our data suggest that this ~5% represents functional elements common to all mammals. The distribution of these elements relative to genes is highly heterogeneous, with roughly half of the most highly conserved non-coding elements apparently devoted to regulating ~1% of human genes; these genes have important roles in development, and understanding the regulatory clusters that surround them may reveal how cellular states are established and maintained. In recent papers^{32,108}, the dog genome sequence has been used to greatly expand the catalogue of mammalian regulatory motifs in promoters and 3'-untranslated regions. The dog genome sequence is also being used to substantially revise the human gene catalogue. Despite these advances, it is clear that mammalian comparative genomics is still in its early stages. Progress will be markedly accelerated by the availability of many additional mammalian genome sequences, initially with light coverage²⁸ but eventually with near-complete coverage.

In addition to its role in studies of mammalian evolution, the dog has a special role in genomic studies because of the unparalleled phenotypic diversity among closely related breeds. The dog is a testament to the power of breeding programmes to select naturally occurring genetic variants with the ability to shape morphology, physiology and behaviour. Genome comparison within and across breeds can reveal the genes that underlie such traits, informing basic research on development and neurobiology. It can also identify disease genes that were carried along in breeding programmes. Potential benefits include insights into disease mechanism, and the possibility of clinical trials in disease-affected dogs to accelerate new therapeutics that would improve health in both dogs and humans. The SNP map of the dog genome confirms that dog breeds show the long-range haplotype structure expected from recent intensive breeding. Moreover, our analysis shows that the current collection of >2.5 million SNPs should be sufficient to allow association studies of

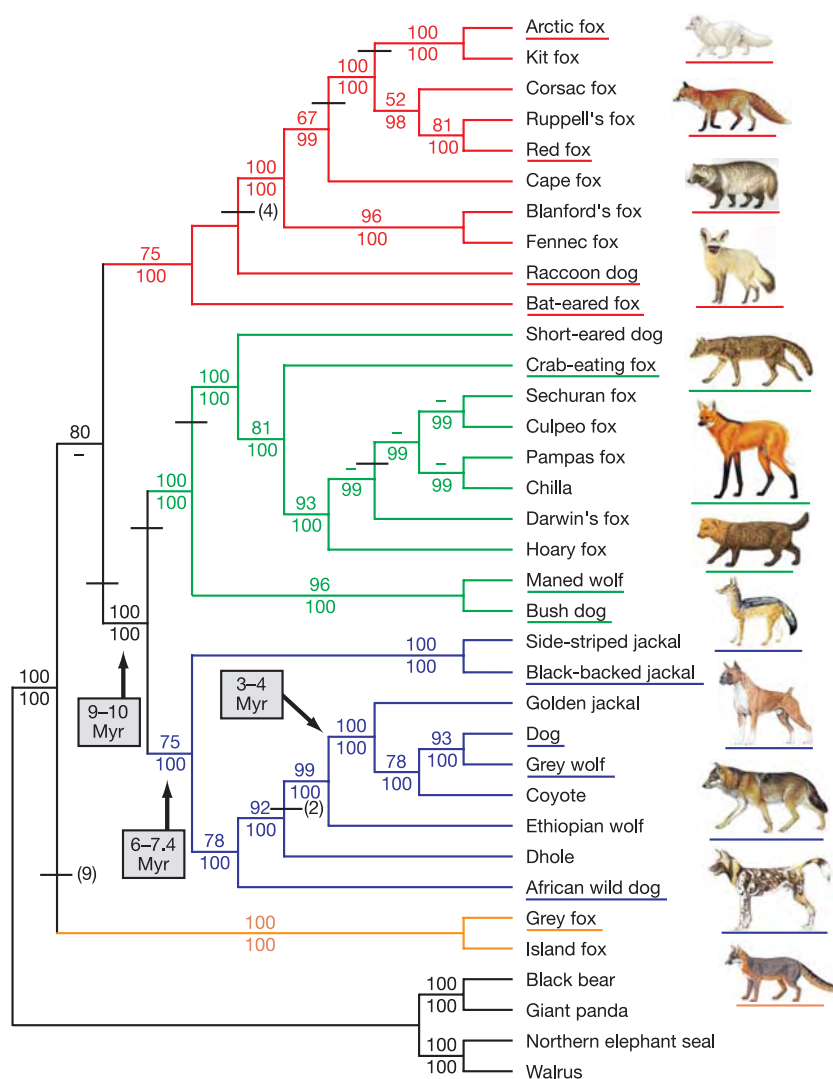


Figure 10 | Phylogeny of canid species. The phylogenetic tree is based on ~15 kb of exon and intron sequence (see text). Branch colours identify the red-fox-like clade (red), the South American clade (green), the wolf-like clade (blue) and the grey and island fox clade (orange). The tree shown was constructed using maximum parsimony as the optimality criterion and is the single most parsimonious tree. Bootstrap values and Bayesian posterior probability values are listed above and below the internodes, respectively; dashes indicate bootstrap values below 50% or Bayesian posterior probability values below 95%. Horizontal bars indicate indels, with the number of indels shown in parentheses if greater than one. Underlined

species names are represented with corresponding illustrations. (Copyright permissions for illustrations are listed in the Supplementary Information.) Divergence time, in millions of years (Myr), is indicated for three nodes as discussed in ref. 1. For scientific names and species descriptions of canids, see ref. 119. A tree based on Bayesian inference differs from the tree shown in two respects: it groups the raccoon dog and bat-eared fox as sister taxa, and groups the grey fox and island fox as basal to the clade containing these sister taxa. However, neither of these topological differences is strongly supported (see text and Supplementary Information).

nearly any trait in any breed. Realizing the full power of dog genetics now awaits the development of appropriate genotyping tools, such as multiplex 'SNP chips'¹⁰⁹—this is already underway. For millennia, dogs have accompanied humans on their travels. It is only fitting that the dog should also be a valued companion on our journeys of scientific discovery.

METHODS

Detailed descriptions of all methods are provided in the Supplementary Information. Links to all of the data can be obtained via the Broad Institute website (<http://www.broad.mit.edu/tools/data.html>).

WGS sequencing and assembly. Approximately 31.5 million sequence reads were derived from both ends of inserts (paired-end reads) from 4-, 10-, 40- and 200-kb clones, all prepared from primary blood lymphocyte DNA from a single female boxer. This particular animal was chosen for sequencing because it had the lowest heterozygosity rate among ~120 dogs tested at a limited set of loci; subsequent analysis showed that the genome-wide heterozygosity rate in this boxer is not substantially different from other breeds⁹¹. The assembly was carried out using an interim version of ARACHNE2+ (<http://www.broad.mit.edu/wga/>).

Genome alignment and comparison. Synteny maps were generated using standard methods²⁴ from pair-wise alignments of repeat masked assemblies using PatternHunter¹¹⁰ on CanFam2.0. All other comparative analyses were performed on BLASTZ/MULTIZ^{111,112} genome-wide alignments obtained from the UCSC genome browser (<http://genome.ucsc.edu>), based on CanFam1.0. Known interspersed repeats were identified and dated using RepeatMasker and DateRepeats¹¹³. The numbers of orthologous nucleotides were counted directly from the alignments using human (hg17) as the reference sequence for all overlaps except the dog–mouse overlap, for which pair-wise (CanFam1.0, mm5) alignments were used.

Divergence rate estimates. Orthologous ancestral repeats were excised from the genome alignment and realigned with the corresponding RepBase consensus using ClustalW. Nucleotide divergence rates were estimated from concatenated repeat alignments using baseml with the REV substitution model¹¹⁴. Orthologous coding regions were excised from the genome alignments using the annotated human coding sequences (CDS) from Ensembl and the UCSC browser Known Genes track (October 2004) as reference. K_A and K_S were estimated for each orthologue triplet using codeml with the F3 × 4 codon frequency model and no additional constraints.

Detection and clustering of sequence conservation. Pair-wise conservation scores and the fraction of orthologous sequences under purifying selection were estimated as in ref. 24. The three-way conservation score S_{HMD} was defined as $S_{HMD} = (p - u) / \sqrt{(u(1 - u)/n)}$, where n is the number of nucleotides aligned across all three genomes (human, mouse, dog) for each non-overlapping 50-bp window with more than 20 aligned bases, p is the fraction of nucleotides identical across all three genomes, and u is the mean identity of ancestral repeats within 500 kb of the window. HCNEs were defined as windows with $S_{HMD} > 5.4$ that did not overlap a coding exon, as defined by the UCSC Known Genes track, and HCNE clusters were defined as all runs of overlapping 1-Mb intervals (50-kb step size) across the human genome with HCNE densities in the 90th percentile.

Gene set acceleration scores. Gene annotation was performed on CanFam1.0. A set of 13,816 orthologous human, mouse and dog genes were identified and compiled into 4,950 gene sets containing genes related by functional annotations or microarray gene expression data. For each gene set S , the acceleration score $A(S)$ along a lineage is defined by (1) ranking all genes based on K_A/K_S within a lineage, (2) calculating the rank-sum statistic for the set along each lineage (denoted $a_{dog}(S)$, $a_{mouse}(S)$, $a_{human}(S)$), (3) calculating the rank-sum for the lineage minus the maximum rank-sum the other lineages, for example, $a_{human}(S) - \max(a_{dog}(S), a_{mouse}(S))$ and (4) converting this rank-sum difference to a z-score by comparing it to the mean and standard deviation observed in 10,000 random sets of the same size. The expected number of sets at a given z-score threshold was estimated by repeating steps (1)–(4) 10,000 times for groups of 4,950 randomly permuted gene sets.

SNP discovery. The SNP discovery was performed on CanFam2.0. Set 1 SNPs were discovered by comparison of the two haplotypes derived from the boxer assembly using only high-quality discrepancies supported by two reads. SNPs in sets 2 and 3 were discovered by aligning reads or contigs to the boxer assembly and using the SSAHA SNP algorithm¹¹⁵.

Haplotype structure. The SNPs within the sequenced boxer genome (CanFam2.0) were assigned to homozygous or heterozygous regions using a Viterbi algorithm¹¹⁶. To determine whether the haplotype structure seen in the boxer is representative of most dog breeds, we randomly selected ten regions of 15 Mb each (~6% of the CanFam2.0 genome) and examined the extent of homozygosity and linkage disequilibrium in these regions in a collection of 224

dogs, consisting of 20 dogs from each of 10 breeds (akita, basenji, bullmastiff, English springer spaniel, Glen of Imaal terrier, golden retriever, Irish wolfhound, Labrador retriever, pug and rottweiler) and one dog from each of 24 additional breeds (see Supplementary Information). For each instance in which a dog was homozygous in a particular 10-kb region, we measured the distance from the beginning of the 10-kb region to the first heterozygous SNP in the adjoining 100-kb, 1-Mb and 15-Mb data. This distance was used as the extent of homozygosity. The boxer sequence was sampled in an identical manner to the actual breed data. Linkage disequilibrium (represented by r^2) across the ten 15-Mb regions was assessed using Haploview¹¹⁷.

Received 9 August; accepted 11 October 2005.

- Wayne, R. K. *et al.* Molecular systematics of the Canidae. *Syst. Biol.* **46**, 622–653 (1997).
- Vila, C. *et al.* Multiple and ancient origins of the domestic dog. *Science* **276**, 1687–1689 (1997).
- Bardeleben, C., Moore, R. L. & Wayne, R. K. Isolation and molecular evolution of the selenocysteine tRNA (*Cf TRSP*) and RNase P RNA (*Cf RPPH1*) genes in the dog family, Canidae. *Mol. Biol. Evol.* **22**, 347–359 (2005).
- Savolainen, P., Zhang, Y. P., Luo, J., Lundeberg, J. & Leitner, T. Genetic evidence for an East Asian origin of domestic dogs. *Science* **298**, 1610–1613 (2002).
- American Kennel Club. *The Complete Dog Book* (eds Crowley, J. & Adelman, B.) (Howell Book House, New York, 1998).
- Wayne, R. K. Limb morphology of domestic and wild canids: the influence of development on morphologic change. *J. Morphol.* **187**, 301–319 (1986).
- Ostrander, E. A., Galibert, F. & Patterson, D. F. Canine genetics comes of age. *Trends Genet.* **16**, 117–123 (2000).
- Patterson, D. Companion animal medicine in the age of medical genetics. *J. Vet. Intern. Med.* **14**, 1–9 (2000).
- Sargan, D. R. IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics. *Mamm. Genome* **15**, 503–506 (2004).
- Chase, K. *et al.* Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc. Natl Acad. Sci. USA* **99**, 9930–9935 (2002).
- Breen, M. *et al.* Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Res.* **11**, 1784–1795 (2001).
- Breen, M., Bullerdiel, J. & Langford, C. F. The DAPI banded karyotype of the domestic dog (*Canis familiaris*) generated using chromosome-specific paint probes. *Chromosome Res.* **7**, 401–406 (1999).
- Breen, M. *et al.* An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics* **5**, 65 (2004).
- Hitte, C. *et al.* Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nature Rev. Genet.* **6**, 643–648 (2005).
- Li, R. *et al.* Construction and characterization of an eightfold redundant dog genomic bacterial artificial chromosome library. *Genomics* **58**, 9–17 (1999).
- Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903 (2003).
- Sutter, N. & Ostrander, E. Dog star rising: The canine genetic system. *Nature Rev. Genet.* **5**, 900–910 (2004).
- Galibert, F., Andre, C. & Hitte, C. Dog as a mammalian genetic model [in French]. *Med. Sci. (Paris)* **20**, 761–766 (2004).
- Pollinger, J. P. *et al.* Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* doi:10.1101/gr.4374505 (in the press).
- Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
- Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
- Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618 (2001).
- Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
- Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102**, 4795–4800 (2005).
- Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).

30. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
31. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3**, e10 (2005).
32. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
33. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
34. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
35. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
36. Richterich, P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* **8**, 251–259 (1998).
37. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5**, R23 (2004).
38. Andelfinger, G. *et al.* Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics* **83**, 1053–1062 (2004).
39. Wang, W. & Kirkness, E. F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res.* doi:10.1101/gr.3765505 (in the press).
40. Mamedov, I. Z., Arzumanyan, E. S., Amosova, A. L., Lebedev, Y. B. & Sverdlov, E. D. Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res.* **33**, e16 (2005).
41. Lin, L. *et al.* The sleep disorder canine narcolepsy is caused by a mutation in the *hypocretin (orexin) receptor 2* gene. *Cell* **98**, 365–376 (1999).
42. Pele, M., Tired, L., Kessler, J. L., Blot, S. & Panthier, J. J. SINE exonic insertion in the *PTPLA* gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum. Mol. Genet.* **14**, 1417–1427 (2005).
43. Fondon, J. W. III & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl Acad. Sci. USA* **101**, 18058–18063 (2004).
44. Galtier, N. & Mouchiroud, D. Isochore evolution in mammals: a human-like ancestral structure. *Genetics* **150**, 1577–1584 (1998).
45. Belle, E. M., Duret, L., Galtier, N. & Eyre-Walker, A. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* **58**, 653–660 (2004).
46. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
47. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
48. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**, 813–820 (2003).
49. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl Acad. Sci. USA* **101**, 13994–14001 (2004).
50. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl Acad. Sci. USA* **90**, 4087–4091 (1993).
51. Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl Acad. Sci. USA* **102**, 140–145 (2005).
52. Laird, C. D., McConaughy, B. L. & McCarthy, B. J. Rate of fixation of nucleotide substitutions in evolution. *Nature* **224**, 149–154 (1969).
53. Li, W. H., Tanimura, M. & Sharp, P. M. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25**, 330–342 (1987).
54. Webber, C. & Ponting, C. P. Hot spots of mutation and breakage in dog and human chromosomes. *Genome Res.* doi:10.1101/gr.3896805 (in the press).
55. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
56. Marques-Bonet, T. & Navarro, A. Chromosomal rearrangements are associated with higher rates of molecular evolution in mammals. *Gene* **353**, 147–154 (2005).
57. Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5**, 15–56 (2004).
58. Smith, N. G., Brandstrom, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84**, 806–813 (2004).
59. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
60. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137–145 (2005).
61. Walter, K., Abnizova, I., Elgar, G. & Gilks, W. R. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet.* **21**, 436–440 (2005).
62. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
63. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
64. Kimura-Yoshida, C. *et al.* Characterization of the pufferfish *Otx2* cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* **131**, 57–71 (2004).
65. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken *Sox2* enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* **4**, 509–519 (2003).
66. de la Calle-Mustienes, E. *et al.* A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate *Iroquois* cluster gene deserts. *Genome Res.* **15**, 1061–1072 (2005).
67. Daly, M. J. Estimating the human gene count. *Cell* **109**, 283–284 (2002).
68. Hogenesch, J. B. *et al.* A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415 (2001).
69. Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12**, 701–709 (2003).
70. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
71. Wolfe, K. H. & Li, W. H. Molecular evolution meets the genomics revolution. *Nature Genet.* **33** (suppl.), 255–265 (2003).
72. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834 (2003).
73. Hughes, A. L. The evolution of the type I interferon gene family in mammals. *J. Mol. Evol.* **41**, 539–548 (1995).
74. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486 (2002).
75. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99**, 16134–16137 (2002).
76. Demetrius, L. Directionality theory and the evolution of body size. *Proc. Biol. Sci.* **267**, 2385–2391 (2000).
77. Fay, J. C. & Wu, C. I. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.* **4**, 213–235 (2003).
78. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101**, 4164–4169 (2004).
79. Mootha, V. K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34**, 267–273 (2003).
80. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
81. Dorus, S. *et al.* Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* **119**, 1027–1040 (2004).
82. Saetre, P. *et al.* From wild wolf to domestic dog: gene expression changes in the brain. *Brain Res. Mol. Brain Res.* **126**, 198–206 (2004).
83. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304–309 (2000).
84. Birkhead, T. R. & Pizzari, T. Postcopulatory sexual selection. *Nature Rev. Genet.* **3**, 262–273 (2002).
85. Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. & Lahn, B. T. Rate of molecular evolution of the seminal protein gene *SEMG2* correlates with levels of female promiscuity. *Nature Genet.* **36**, 1326–1329 (2004).
86. Ruiz-Pesini, E. *et al.* Correlation of sperm motility with mitochondrial enzymatic activities. *Clin. Chem.* **44**, 1616–1620 (1998).
87. Zeh, J. A. & Zeh, D. W. Maternal inheritance, sexual conflict and the maladapted male. *Trends Genet.* **21**, 281–286 (2005).
88. Grossman, L. I., Wildman, D. E., Schmidt, T. R. & Goodman, M. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends Genet.* **20**, 578–585 (2004).
89. Ostrander, E. A. & Kruglyak, L. Unleashing the canine genome. *Genome Res.* **10**, 1271–1274 (2000).
90. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **12**, 2388–2396 (2004).
91. Parker, H. G. *et al.* Genetic structure of the purebred domestic dog. *Science* **304**, 1160–1164 (2004).
92. Bardeleben, C., Moore, R. L. & Wayne, R. K. A molecular phylogeny of the Canidae based on six nuclear loci. *Mol. Phylogenet. Evol.* **37**, 815–831 (2005).
93. Fogel, B. *The Encyclopedia of the Dog* (D.K. Publishing, New York, 1995).
94. Wilcox, B. & Walkowicz, C. *The Atlas of Dog Breeds of the World* (T.H.F. Publications, Neptune City, New York, 1995).
95. Frazer, K. A. *et al.* Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res.* **14**, 1493–1500 (2004).
96. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* Vol. 7 (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).
97. Vila, C., Seddon, J. & Ellegren, H. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet.* **21**, 214–218 (2005).

98. Leonard, J. A. *et al.* Ancient DNA evidence for Old World origin of New World dogs. *Science* **298**, 1613–1616 (2002).
99. Kajiwar, N. & Japanese Kennel Club in Akita (eds Kariyabu, T. & Kaluzniacki, S.) 1–103 (Japan Kennel Club, Tokyo, 1998).
100. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
101. Werner, P., Raducha, M. G., Prociuk, U., Henthorn, P. S. & Patterson, D. F. Physical and linkage mapping of human chromosome 17 loci to dog chromosomes 9 and 5. *Genomics* **42**, 74–82 (1997).
102. Todhunter, R. J. *et al.* Power of a Labrador Retriever-Greyhound pedigree for linkage analysis of hip dysplasia and osteoarthritis. *Am. J. Vet. Res.* **64**, 418–424 (2003).
103. Sidjanin, D. J. *et al.* Canine *CNGB3* mutations establish cone degeneration as orthologous to the human achromatopsia locus *ACHM3*. *Hum. Mol. Genet.* **11**, 1823–1833 (2002).
104. Lou, X. Y. *et al.* The extent and distribution of linkage disequilibrium in a multi-hierarchical outbred canine pedigree. *Mamm. Genome* **14**, 555–564 (2003).
105. Hyun, C. *et al.* Prospects for whole genome linkage disequilibrium mapping in domestic dog breeds. *Mamm. Genome* **14**, 640–649 (2003).
106. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19**, 135–140 (2003).
107. Tsui, C. *et al.* Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res.* **31**, 4910–4916 (2003).
108. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
109. Syvanen, A. C. Toward genome-wide SNP genotyping. *Nature Genet.* **37** (suppl.), S5–10 (2005).
110. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).
111. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
112. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
113. Smit, A. F. A. & Green, P. RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>).
114. Yang, Z., Goldman, N. & Friday, A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**, 316–324 (1994).
115. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
116. Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Inform. Process.* **13**, 260–269 (1967).
117. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
118. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
119. Macdonald, D. W. & Sillero-Zubiri, C. in *Biology and Conservation of Canids* (eds Macdonald, D. W. & Sillero-Zubiri, C.) 1–30 (Oxford Univ. Press, Oxford, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to the canine research community, and in particular D. Patterson, G. Acland and K. G. Lark, whose vision and research convinced the NIH of the importance of generating a canine genome sequence. We also thank all those who shared insights at the Dog Genome Community meetings, including G. Acland, G. D. Aguirre, M. Binns, U. Giger, P. Henthorn, F. Lingaas, K. Murphy and P. Werner. We thank our many colleagues (G. Acland, G. D. Aguirre, C. Andre, N. Fretwell, G. Johnson, K. G. Lark and J. Modiano), as well as the dog owners and breeders who provided us with samples. We thank colleagues at the UCSC browser for providing data (such as BLASTZ alignments), A. Smit for providing the RepeatMasker annotations used in our analyses and N. Manoukis for providing Unix machines for the phylogenetic analyses. Finally, we thank L. Gaffney and K. Siang Toh for editorial and graphical assistance. The genome sequence and analysis was supported in part by the National Human Genome Research Institute. The radiation hybrid map was supported in part by the Canine Health Foundation. Sample collection was supported in part by the Intramural Research Program of the National Human Genome Research Institute and the Canine Health Foundation.

Author Information The draft genome sequence has been deposited in public databases under NCBI accession codes AAEX01000000 (CanFam1.0) and AAEX02000000 (CanFam2.0). SNPs have been deposited in the dbSNP database (<http://www.ncbi.nlm.nih.gov/projects/SNP/>). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.L.T. (kersli@broad.mit.edu) or E.S.L. (lander@broad.mit.edu).

Broad Sequencing Platform members Jennifer Baldwin¹, Adal Abebe¹, Amr Abouelleil¹, Lynne Aftuck¹, Mostafa Ait-zahra¹, Tyler Aldredge¹, Nicole Allen¹, Peter An¹, Scott Anderson¹, Claudel Antoine¹, Harindra Arachchi¹, Ali Aslam¹, Laura Ayotte¹, Pasang Bachantsang¹, Andrew Barry¹, Tashi Bayul¹, Mostafa Benamara¹, Aaron Berlin¹, Daniel Besette¹, Berta Blitshteyn¹, Toby Bloom¹, Jason Blye¹, Leonid Boguslavskiy¹, Claude Bonnet¹, Boris Boukhgalter¹, Adam Brown¹, Patrick Cahill¹, Nadia Calixte¹, Jody Camarata¹, Yama Cheshatsang¹, Jeffrey Chu¹, Mieke Citroen¹, Alville Collymore¹, Patrick Cooke¹, Tenzin Dawoe¹, Riza Daza¹, Karin Decktor¹, Stuart DeGray¹, Norbu Dhargay¹, Kimberly Dooley¹, Kathleen Dooley¹, Passang Dorje¹, Kunsang Dorjee¹, Lester Dorris¹, Noah Duffey¹, Alan Dupes¹, Osebhajajeme Egbiremolun¹, Richard Elong¹, Jill Falk¹, Abderrahim Farina¹, Susan Faro¹, Diallo Ferguson¹, Patricia Ferreira¹, Sheila Fisher¹, Mike FitzGerald¹, Karen Foley¹, Chelsea Foley¹, Alicia Franke¹, Dennis Friedrich¹, Diane Gage¹, Manuel Garber¹, Gary Gearin¹, Georgia Giannoukos¹, Tina Goode¹, Audra Goyette¹, Joseph Graham¹, Edward Grandbois¹, Kunsang Gyaltzen¹, Nabil Hafez¹, Daniel Hagopian¹, Birhane Hagos¹, Jennifer Hall¹, Claire Healy¹, Ryan Hegarty¹, Tracey Honan¹, Andrea Horn¹, Nathan Houde¹, Leanne Hughes¹, Leigh Hunnicutt¹, M. Husby¹, Benjamin Jester¹, Charlien Jones¹, Asha Kamat¹, Ben Kanga¹, Cristyn Kells¹, Dmitry Khazanovich¹, Alix Chinh Kieu¹, Peter Kisner¹, Mayank Kumar¹, Krista Lance¹, Thomas Landers¹, Marcia Lara¹, William Lee¹, Jean-Pierre Leger¹, Niall Lennon¹, Lisa Leuper¹, Sarah LeVine¹, Jinlei Liu¹, Xiaohong Liu¹, Yeshi Lokyitsang¹, Tashi Lokyitsang¹, Annie Lui¹, Jan Macdonald¹, John Major¹, Richard Marabella¹, Kebede Maru¹, Charles Matthews¹, Susan McDonough¹, Teena Mehta¹, James Meldrim¹, Alexandre Melnikov¹, Louis Meneus¹, Atanas Mihalev¹, Tanya Mihova¹, Karen Miller¹, Rachel Mittelman¹, Valentine Mlenga¹, Leonidas Mulrain¹, Glen Munson¹, Adam Navidi¹, Jerome Naylor¹, Tuyen Nguyen¹, Nga Nguyen¹, Cindy Nguyen¹, Thu Nguyen¹, Robert Nicol¹, Nyima Norbu¹, Choe Norbu¹, Nathaniel Novod¹, Tenchoe Nyima¹, Peter Olandt¹, Barry O'Neill¹, Keith O'Neill¹, Sahal Osman¹, Lucien Oyono¹, Christopher Patti¹, Danielle Perrin¹, Pema Phunkhang¹, Fritz Pierre¹, Margaret Priest¹, Anthony Rachupka¹, Sujaa Raghuraman¹, Rayale Rameau¹, Verneda Ray¹, Christina Raymond¹, Filip Rege¹, Cecil Rise¹, Julie Rogers¹, Peter Rogov¹, Julie Sahalie¹, Sampath Settipalli¹, Theodore Sharpe¹, Terrance Shea¹, Mechele Sheehan¹, Ngawang Sherpa¹, Jianying Shi¹, Diana Shih¹, Jessie Sloan¹, Cherylyn Smith¹, Todd Sparrow¹, John Stalker¹, Nicole Stange-Thomann¹, Sharon Stavropoulos¹, Catherine Stone¹, Sabrina Stone¹, Sean Sykes¹, Pierre Tchuinga¹, Pema Tenzing¹, Senait Tesfaye¹, Dawa Thoulutsang¹, Yama Thoulutsang¹, Kerri Topham¹, Ira Topping¹, Tsamla Tsamla¹, Helen Vassiliev¹, Vijay Venkataraman¹, Andy Vo¹, Tsering Wangchuk¹, Tsering Wangdi¹, Michael Weiland¹, Jane Wilkinson¹, Adam Wilson¹, Shailendra Yadav¹, Shuli Yang¹, Xiaoping Yang¹, Geneva Young¹, Qing Yu¹, Joanne Zainoun¹, Lisa Zembek¹ & Andrew Zimmer¹

ARTICLES

VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche

Rosandra N. Kaplan^{1,2,6*}, Rebecca D. Riba^{1,2*}, Stergios Zacharoulis^{1,2,6*}, Anna H. Bramley^{1,2}, Loïc Vincent⁴, Carla Costa^{1,2}, Daniel D. MacDonald^{1,2}, David K. Jin⁴, Koji Shido⁴, Scott A. Kerns^{1,2}, Zhenping Zhu⁸, Daniel Hicklin⁸, Yan Wu⁸, Jeffrey L. Port⁵, Nasser Altorki⁵, Elisa R. Port⁷, Davide Ruggero⁹, Sergey V. Shmelkov^{1,2,4}, Kristian K. Jensen^{1,2}, Shahin Rafii^{3,4} & David Lyden^{1,2,6}

The cellular and molecular mechanisms by which a tumour cell undergoes metastasis to a predetermined location are largely unknown. Here we demonstrate that bone marrow-derived haematopoietic progenitor cells that express vascular endothelial growth factor receptor 1 (VEGFR1; also known as Flt1) home to tumour-specific pre-metastatic sites and form cellular clusters before the arrival of tumour cells. Preventing VEGFR1 function using antibodies or by the removal of VEGFR1⁺ cells from the bone marrow of wild-type mice abrogates the formation of these pre-metastatic clusters and prevents tumour metastasis, whereas reconstitution with selected Id3 (inhibitor of differentiation 3)-competent VEGFR1⁺ cells establishes cluster formation and tumour metastasis in Id3 knockout mice. We also show that VEGFR1⁺ cells express VLA-4 (also known as integrin $\alpha_4\beta_1$), and that tumour-specific growth factors upregulate fibronectin—a VLA-4 ligand—in resident fibroblasts, providing a permissive niche for incoming tumour cells. Conditioned media obtained from distinct tumour types with unique patterns of metastatic spread redirected fibronectin expression and cluster formation, thereby transforming the metastatic profile. These findings demonstrate a requirement for VEGFR1⁺ haematopoietic progenitors in the regulation of metastasis, and suggest that expression patterns of fibronectin and VEGFR1⁺VLA-4⁺ clusters dictate organ-specific tumour spread.

Bone marrow-derived cells (BMDCs) contribute to malignant transformation¹, tumour vascularization^{2,3} and neoplastic cell migration⁴. Previously, we identified haematopoietic progenitor cells (HPCs) expressing VEGFR1 that reside within specified niches of the bone marrow. During the angiogenic switch, these cells proliferate and mobilize to the bloodstream along with bone marrow-derived endothelial progenitor cells that express VEGFR2 (also known as Flk1), and contribute to the vascularization and growth of specific primary tumours^{2,5}. These myelomonocytic VEGFR1⁺ cells localize to perivascular sites, thus stabilizing tumour neo-vessels². These and other tumour-associated cells enhance primary tumour neo-angiogenesis and growth, yet their precise contribution to metastasis is unclear^{6–8}. Therefore, the aim of this study was to determine the role of VEGFR1⁺ HPCs in the temporal and functional generation of metastasis.

BMDCs colonize pre-metastatic sites before tumour cells

We analysed the fate of β -galactosidase-positive (β -gal⁺) and green fluorescent protein-positive (GFP⁺) BMDCs following intradermal primary tumour injection in mice. Animals were inoculated with either Lewis lung carcinoma (LLC) cells, which metastasize to the lungs and occasionally the liver, or B16 melanoma cells, which possess a more widely disseminated metastatic potential. After

irradiation, but before tumour implantation, we observed minimal β -gal⁺ BMDCs (mean $\pm$ s.e.m., 0.01% $\pm$ 0.01 of cells β -gal⁺ per $\times 100$ objective field) or GFP⁺ BMDCs in the lungs (Fig. 1a, b, left panels). By day 14 after tumour implantation, but before the arrival of tumour cells, the extravasation and cluster formation of β -gal⁺ BMDCs (3.2% $\pm$ 1.2, $P < 0.05$ by Student's *t*-test) or GFP⁺ BMDCs were detected near terminal bronchioles and distal alveoli, both common sites for future tumour metastasis (Fig. 1a, b, left middle panels and insets). On day 16, established β -gal⁺ cell clusters dictated the contours of future metastatic lesions (Fig. 1a, right middle panel). Individual DsRed-tagged tumour cells, associated with pre-existing BMDC clusters, were visible by day 18 (Fig. 1b, right middle panel) and progressed to micrometastases by day 23 (Fig. 1a, b, right panels). β -gal⁺ BMDCs were maintained within well-established tumour metastases (Fig. 1a, right panel and inset).

To further define the timing of tumour cell arrival, a flow cytometric study of the lungs was undertaken. Before day 8, minimal GFP⁺ BMDCs were observed in this tissue; however, from day 12, BMDCs began migrating into the lung (Fig. 1c, graph and left flow cytometry panel). These GFP⁺ cells increased in number, and were joined by DsRed-tagged tumour cells by day 18 (Fig. 1c, graph and right flow diagram). No tumour cells were detected by flow cytometry or microscopy earlier than day 16, and increasing numbers

¹Department of Pediatrics and the Children's Blood Foundation Laboratories, ²Cell and Developmental Biology, ³Howard Hughes Medical Institute, ⁴Genetic Medicine and ⁵Surgery, Weill Cornell Medical College of Cornell University, 1300 York Avenue and ⁶Department of Pediatrics and ⁷Surgery, Memorial Sloan-Kettering Cancer Center, 1233 York Avenue, New York, New York 10021, USA. ⁸Imclone Systems Incorporated, New York, New York 10014, USA. ⁹Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111, USA.

*These authors contributed equally to this work.

of tumour cells were identified over time (Fig. 1b, right panels; Fig. 1c; Supplementary Fig. 1a). More than 95% of tumour cells co-clustered with GFP⁺ BMDCs (97% $\pm$ 1.1; Fig. 1b, right panels).

Although a few tumour cells may have been undetectable using these methodologies, further experiments in mice given B16-melanoma-conditioned media (MCM) showed that this conditioning alone mobilized BMDCs that were capable of forming a pre-metastatic niche. We introduced DsRed-tagged B16 tumour cells intravenously into mice with pre-established GFP⁺ BMDC clusters in the lung after challenge with MCM (Fig. 1d, right panel) or media alone (Fig. 1d, left panel). MCM increased the number of tumour cells in the lung one day after tumour injection compared with media alone (141.3 $\pm$ 10.2 versus 2.7 $\pm$ 0.6 tumour cells per section of lung tissue, P < 0.01). Four days after tumour injection, the frequency

and size of the lung nodules were augmented by MCM (207 $\pm$ 5.6 versus 14 $\pm$ 1.7, P < 0.01; Fig. 1d, right panel inset). Co-localization of DsRed-tagged tumour cells with GFP⁺ BMDC clusters was >93% at both time points, indicating that BMDCs assist tumour cell adhesion and proliferation. Therefore, factors provided by the primary tumour induce BMDCs to enter the bloodstream and mobilize to organ-specific pre-metastatic sites, and this migration precedes the arrival of tumour cells.

Sites of BMDC clusters are tumour-type specific

We examined whether the type of tumour cell dictated BMDC distribution to specific pre-metastatic sites. Intradermal injection of LLC cells resulted in BMDC cluster formation limited to the lung (47.5 $\pm$ 2.6 clusters per $\times$ 100 objective field) and liver (10.8 $\pm$ 1.1)

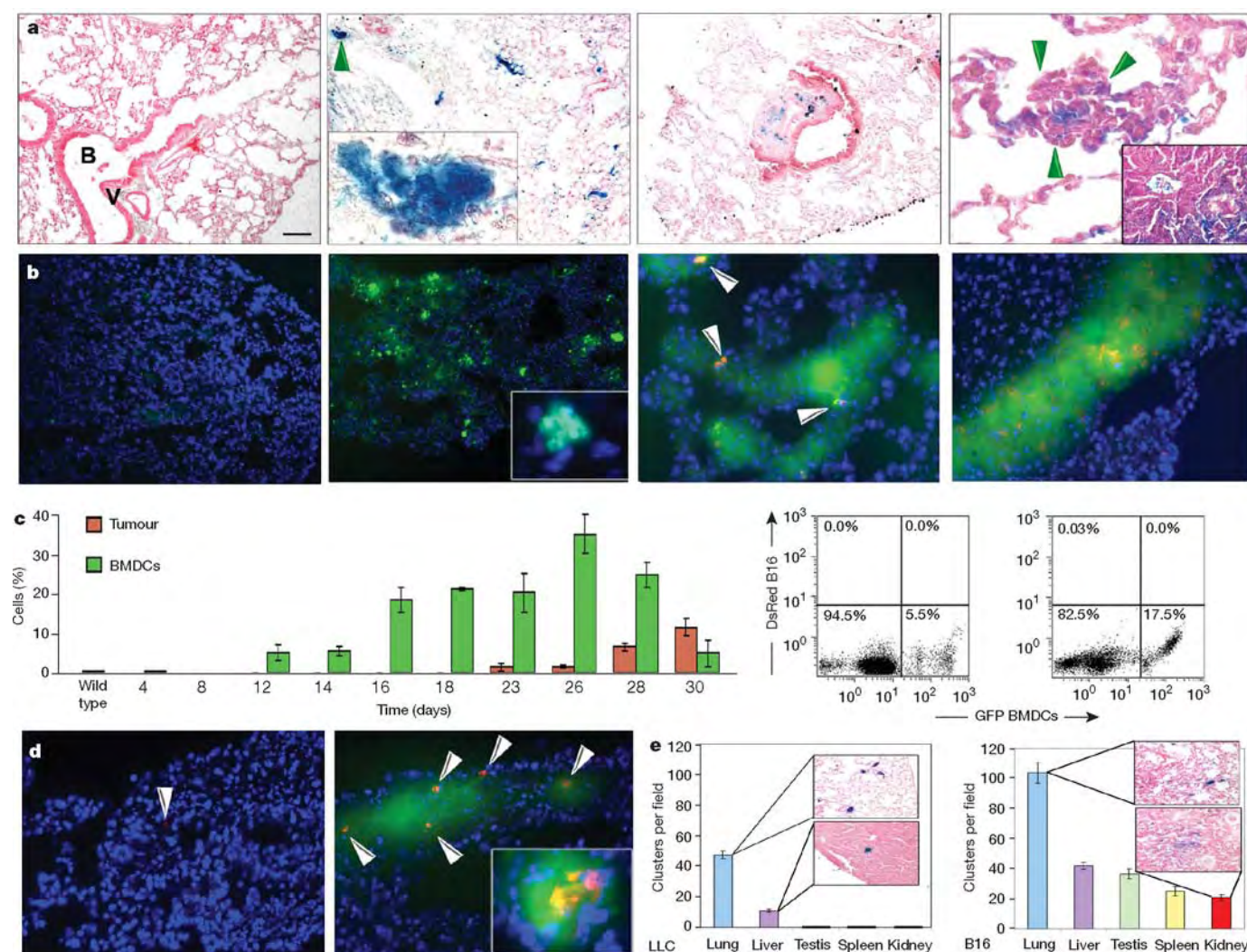


Figure 1 | Bone marrow-derived cells form the pre-metastatic niche.

a, β -gal⁺ bone marrow cells (left panel) are rarely observed in lungs after irradiation and before LLC cell implantation (n = 6). By day 14, β -gal⁺ bone marrow-derived clusters appear in the lung parenchyma (left middle panel and magnified inset of the region arrowed; n = 25) and are associated with micrometastases by day 23 (right panel, arrows) and in gross metastases (right panel, inset; n = 12). Also shown is a cluster with associated stroma between a terminal bronchiole and bronchial vein, a common metastatic site (right middle panel). B, terminal bronchiole; V, bronchial vein. **b**, GFP⁺ bone marrow in the lungs after irradiation and before DsRed-tagged B16 cell implantation (left panel; n = 6). On day 14, GFP⁺ (green) BMDCs are seen with no DsRed⁺ (red) tumour cells (left middle panel and inset; n = 12). Beginning on day 18, a few single DsRed⁺ B16 cells adhere to GFP⁺ bone marrow clusters (right middle panel), and by day 23, DsRed⁺ tumour cells

proliferate at cluster sites (right panel; n = 8). DAPI stain (blue) shows cell nuclei. **c**, A graph showing flow cytometric data of bone marrow-derived GFP⁺ BMDCs and DsRed⁺ B16 cells in the lung, and two flow diagrams on day 14 (left panel) and day 18 (right panel) (n = 30; error bars show s.e.m.). **d**, GFP⁺ BMDCs mobilized by B16 conditioned media, then DsRed-tagged tumour cells injected through the tail vein adhere 24 h later (right panel, arrows) compared with animals receiving media alone (left panel; P < 0.01). Inset shows proliferating tumour cells in a cluster after four days (right panel inset; n = 6). **e**, Number of clusters per $\times$ 100 objective field in animals with intradermal LLC or B16 tumours (n = 12). Scale bar on top left panel applies to panels **a** (left, left middle, right middle, 80 μ m; left middle inset, 8 μ m; right, 20 μ m; right inset, 47 μ m), **b** (left, left middle, 80 μ m; left middle inset, 8 μ m; right middle, right, 40 μ m) and **d** (40 μ m; right inset, 20 μ m).

with no clusters in other organs (Fig. 1e, left panel). In contrast, the B16 melanoma tumour cells induced the formation of BMDC clusters in multiple tissues such as the lung (103.8 ± 6.9), liver (41.8 ± 2.4), testis (36.6 ± 3.1), spleen (25 ± 3.2) and kidney (20.6 ± 1.8), which are all common metastatic sites for this tumour (Fig. 1e, right panel). Furthermore, melanoma cells, consistent with their more aggressive metastatic nature, induced more clusters than LLC cells ($P < 0.01$).

Recruited BMDCs consist of haematopoietic progenitors

We characterized the cellular and molecular composition of incorporated BMDC clusters. Clusters induced by either tumour type expressed VEGFR1 (Fig. 2a, right panel), and GFP⁺ BMDC clusters coexpressed VEGFR1 (Fig. 2b, left panel), compared with little VEGFR1 in the lung after irradiation alone (Fig. 2a, left panel and inset). Further characterization revealed that subsets of VEGFR1⁺ BMDCs coexpressed the stem/progenitor cell antigens CD133 (Fig. 2b, right panel), CD34 (Supplementary Fig. 1b and Supplementary Table) and CD117 (also known as c-Kit; Fig. 2c), suggesting that these cells may comprise phenotypically marked VEGFR1⁺ HPCs and precursor cells. After primary tumour implantation, CD117-positive progenitor cells arrived in the lung before GFP-tagged tumour cells by flow cytometry (Supplementary Fig. 1c), recapitulating the recruitment of BMDCs described above. There is a degree of maturational heterogeneity, with the myelomonocytic marker CD11b present on certain incorporated cells (data not shown). Early VEGFR1⁺ bone marrow clusters lacked expression of VEGFR2 and CD31 (also known as PECAM1; Supplementary Fig. 1d, left and left middle panels, respectively). VEGFR2-positive circulating endothelial progenitor cells migrated to fully formed BMDC clusters (Supplementary Fig. 1d, right panel), and coincided with the arrival of tumour cells (Supplementary Fig. 1e, graph). Thus bone marrow-derived VEGFR1⁺ HPCs initiate and maintain the pre-metastatic niche.

BMDC clusters occur in a spontaneous tumour model

We compared these findings to those in a spontaneous tumour model using c-Myc transgenic mice. On day 40 of life, prominent VEGFR1⁺

clusters were detected exclusively in the lymph nodes of these animals before the onset of lymphoma (145.1 ± 16.4 clusters per $\times 100$ objective field; Fig. 2d, middle panel and inset), with no observed clusters in wild-type littermates (0.4 ± 0.3 , $P < 0.001$; Fig. 2d, left panel). By 120 days, VEGFR1⁺ clusters persisted in established lymphomas (67.8 ± 9.5 versus 0.7 ± 0.5 in c-Myc mice versus littermates, $P < 0.001$; Fig. 2d, right panel and inset). The lymphoma cells, which surrounded the VEGFR1⁺ HPCs, did not express VEGFR1 (Fig. 2d, right panel inset).

BMDC clusters are recruited to pre-metastatic human tissue

To validate the mouse data showing tumour-specific formation of VEGFR1⁺ cellular clusters, we analysed human tissues from patients with malignancy. VEGFR1⁺ clusters were observed in both primary tumours and metastatic tissue (Fig. 3, showing breast carcinoma in an axillary lymph node, lung carcinoma and oesophageal carcinoma). There were increased cellular clusters in common sites of metastasis before tumour spread, suggesting the potential of this tissue as a future site for metastasis (Fig. 3, showing axillary lymph node (21 ± 5 clusters per $\times 100$ objective field), lung (19 ± 4) and gastro-oesophageal junction (25 ± 4)). In patients without malignancy, lymph nodes and lung tissue did not show VEGFR1⁺ clusters (Fig. 3b, d, insets). VEGFR1⁺ cellular clusters expressed the haematopoietic progenitor marker c-Kit (Fig. 3e, f, insets).

Functional role for VEGFR1⁺ BMDCs in directing metastasis

We assessed the potential of purified VEGFR1⁺ bone marrow cells to initiate pre-metastatic clusters by selectively transplanting these progenitors into irradiated mice. By day 24 after LLC tumour cell implantation, control mice that received wild-type bone marrow showed prominent lung metastases and established blood vessels (Fig. 4a, left panel and inset). However, mice transplanted with purified VEGFR1⁺ cells formed numerous micrometastases throughout the lungs (25 ± 9 micrometastases per $\times 100$ objective field; Fig. 4a, middle panel) with aberrant vasculature (Fig. 4a, middle panel inset). In contrast, bone marrow depleted of VEGFR1⁺ cells failed to produce pre-metastatic clusters (Fig. 4a, right panel; $P < 0.01$ by analysis of variance (ANOVA)). These results suggest

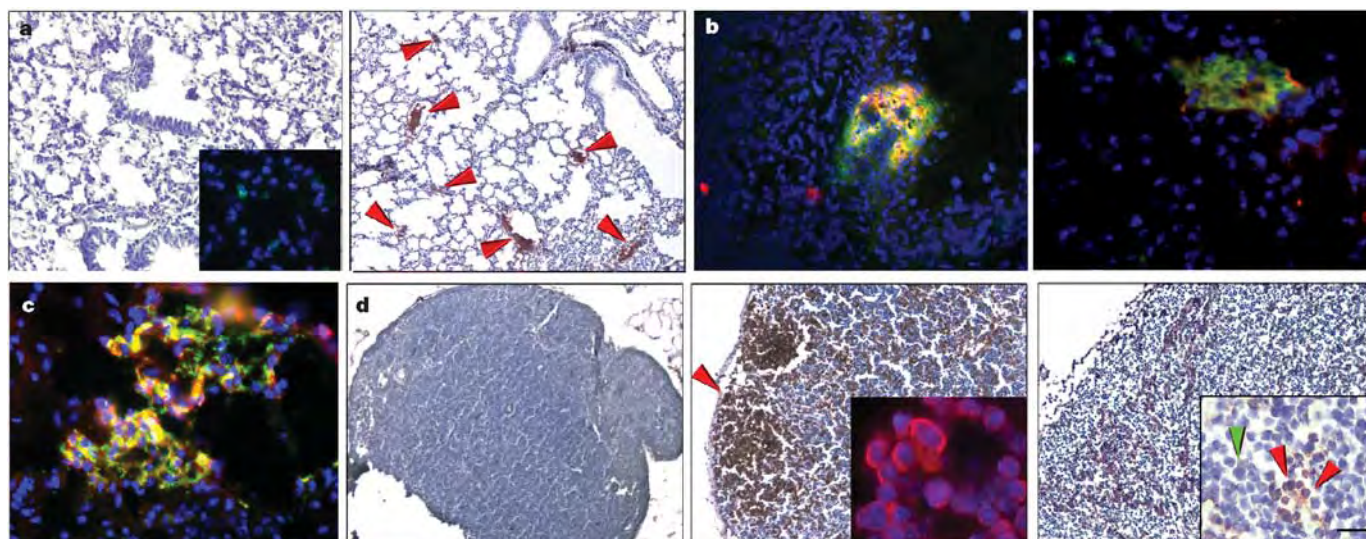


Figure 2 | Pre-metastatic clusters are comprised of VEGFR1⁺ haematopoietic progenitors. **a**, VEGFR1 staining in irradiated lung before tumour implantation (left panel and inset; $n = 10$) and 14 days after LLC cell implantation showing clusters in the lung (right panel, arrows; $n = 18$, $3.9 \pm 0.2\%$ cells with VEGFR1 staining per $\times 100$ objective field, $P < 0.05$). **b**, **c**, Double immunofluorescence in the lung of an animal with day 14 LLC tumour. **b**, VEGFR1⁺ (red) and GFP⁺ (green) bone marrow cells (left panel), VEGFR1⁺ (red) and CD133⁺ (green) (right panel). **c**, VEGFR1⁺

(red) and CD117⁺ (green). **d**, VEGFR1⁺ clusters in c-Myc transgenic lymph node at day 40 of life and before tumorigenesis (middle panel and inset showing VEGFR1⁺ cells (red)) as compared with wild-type littermate lymph node without the transgene (left panel), and day 120 c-Myc transgenic node with lymphoma (right panel). In the inset of the right panel, arrows indicate the VEGFR1⁺ clusters (red) surrounded by lymphoma (green) ($n = 6$). Scale bar at bottom right applies to panels **a** (80 μm ; left inset, 40 μm), **b** (20 μm), **c** (20 μm) and **d** (80 μm ; insets, 8 μm).

that the VEGFR1⁺ HPCs initiating the pre-metastatic cluster can attract tumour cells.

To address whether disruption of VEGFR1⁺ cellular cluster formation could block the metastasis of well-established tumours, mice inoculated with LLC or B16 tumour cells were treated with monoclonal antibodies against VEGFR1 and/or VEGFR2. This approach allows for selective targeting of the BMDCs, as the tumour cells do not express either VEGFR1 or VEGFR2. By day 24, widespread metastases were evident in untreated mice with LLC tumours in the lung (Fig. 4b, left panel) or B16 tumours in the spleen (Supplementary Fig. 2, left panel and inset). Anti-VEGFR1 antibody treatment eliminated the initiating clusters and completely prevented metastasis (Fig. 4b, left middle panel; Supplementary Fig. 3; $P < 0.01$ by ANOVA), whereas anti-VEGFR2 antibody did not prevent the formation of VEGFR1⁺ clusters but limited metastatic progression (15 ± 11 micrometastases per $\times 100$ objective field; Fig. 4b, right middle panel and inset; Supplementary Fig. 3). The two antibodies combined blocked cluster formation to an extent similar to anti-VEGFR1 therapy; however, we did observe an isolated LLC lesion in the lung of one animal (Supplementary Fig. 3b, inset). Collectively, these results suggest that targeting the VEGFR1⁺ cell cluster can prevent tumour cell adhesion, proliferation and metastatic spread.

VLA-4, MMP9 and Id3 mediate the pre-metastatic niche

We investigated the cellular and molecular mechanisms by which migratory HPCs, through interaction with the microenvironment,

form permissive pre-metastatic niches. The interaction of VLA-4 (integrin $\alpha_4\beta_1$) with its ligand fibronectin is essential for the migration of haematopoietic cells within the bone marrow^{9,10} and of circulating leukocytes^{4,11}. We assessed whether VEGFR1⁺ cells express integrins, which may facilitate the interaction of this cell type with the pre-metastatic niche. We found that VEGFR1⁺ HPCs at the pre-metastatic cluster express VLA-4 (Fig. 5a, and inset showing coexpression with VEGFR1), suggesting that VLA-4 allows for the adhesion of the BMDCs that form the pre-metastatic niche. Following cluster formation, $\alpha_4\beta_7$ and $\alpha_6\beta_4$ integrins were prominently expressed within the metastatic niche (data not shown). Proteinases including matrix metalloproteinase 9 (MMP9), produced by haematopoietic cells, can serve to break down basement membranes, thus altering the local microenvironments by releasing soluble Kit-ligand and VEGF-A to support newly introduced cells that express c-Kit^{12,13}. In addition, metalloproteinase expression can be enhanced through $\alpha_4\beta_1$ signalling after fibronectin binding^{14,15}. MMP9 was expressed in pre-metastatic clusters, and this upregulation of MMP9 expression may be a result of integrin binding and activation in VEGFR1⁺ HPCs (Fig. 5b). These findings expand upon previous work demonstrating that VEGFR1-mediated induction of MMP9 directed metastasis to the lungs⁷.

We previously showed that upregulation of Id gene expression is critical for the mobilization of progenitors that aid the growth of primary tumours¹⁵. Id3 expression was also seen within the clusters (Fig. 5c, and inset showing coexpression with VEGFR1). Id3 may facilitate the mobilization of VEGFR1⁺ cells to the pre-metastatic

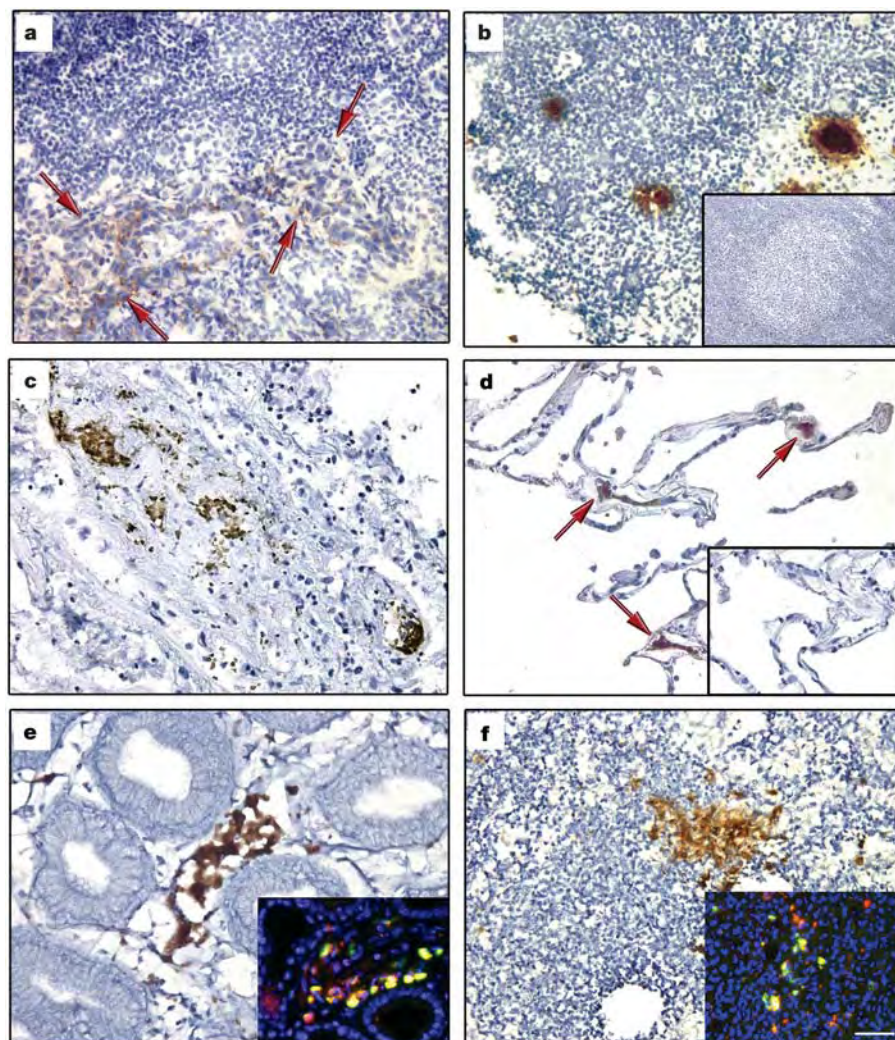


Figure 3 | Expression of VEGFR1 in pre-metastatic human tissue. **a–f**, Cellular clusters stained with VEGFR1 in malignant and non-malignant tissues in individuals with breast ($n = 15$), lung ($n = 15$) and gastrointestinal ($n = 3$) cancers. Lymph node with evidence of breast adenocarcinoma metastasis (**a**, red arrows indicate tumour) and lymph node without malignancy from same patient (**b**). Primary lung adenocarcinoma (**c**) and adjacent 'normal' lung without neoplasm (**d**, red arrows indicate VEGFR1⁺ cells). No VEGFR1⁺ clusters were seen in lymph node (**b**, inset; $n = 6$) and lung tissue (**d**, inset; $n = 3$) from individuals without cancer. Also shown is a primary adenosquamous carcinoma of the gastro-oesophageal junction (**e**), and a hepatic lymph node without carcinoma (**f**). Insets in **e**, **f**, show co-immunofluorescence of VEGFR1 (red) and c-Kit (green). Scale bar at bottom right applies to all panels (40 μm ; insets, 40 μm).

niche. In addition, expression of specific integrins is regulated by Id genes, and may be responsible for BMDC and stromal cell interactions, motility and recruitment¹⁶.

To confirm the functional roles of these proteins in establishing the pre-metastatic niche, we either inhibited the expression of VLA-4 (with anti-integrin $\alpha 4$ antibodies) or studied VEGFR1⁺ cell cluster formation in MMP9 and Id3 knockout mice. In these models, we found reduced cluster formation (Supplementary Fig. 3a–c) and metastatic spread three weeks after tumour implantation. We also found impaired mobilization of VEGFR1⁺ HPCs into the circulation of Id3 knockout mice compared to wild type (654 versus 3,283 VEGFR1⁺CD11b⁺ cells μl^{-1}) in response to tumour inoculation ($P < 0.01$ by Student's *t*-test; Supplementary Table). Decreased mobilization of HPCs may explain the reduced metastatic phenotype seen in these animals^{2,17}.

To formally examine the potential of wild-type VEGFR1⁺ cells to restore the metastatic defect in Id3 knockout mice, Id3-competent GFP⁺VEGFR1⁺ HPCs were injected intravenously into Id3 knockout tumour-bearing mice. VEGFR1⁺ HPCs alone re-established cluster formation and micrometastases by day 21 after tumour implantation (Fig. 5d, and upper inset; Supplementary Fig. 3c). Notably, the LLC metastatic lesions were associated with GFP⁺ BMDCs (Fig. 5d, lower inset). These findings further emphasize the functional role of VEGFR1⁺ BMDCs in the establishment of clusters and metastasis.

Fibronectin upregulation supports adhesion of VLA-4⁺ BMDCs

We next investigated the potential of tissue-specific ligands to support the adhesion and formation of BMDC clusters. Following the implantation of LLC tumour cells, but before the homing of the VLA-4⁺VEGFR1⁺ BMDCs, increased fibronectin expression was observed from day 3 (Fig. 5e, middle panel; Fig. 5f) to day 14 (Fig. 5e, right panel; Fig. 5f) in the vicinity of the future metastatic niche, compared with the baseline level of fibronectin expression in wild-type lung (Fig. 5e, left panel; Fig. 5f). Furthermore, resident fibroblast-like stromal cells (Fig. 5e, left panel inset), which proliferate in response to primary tumour (Fig. 5e, right panel inset), may

contribute to the localized deposition of fibronectin. Melanoma cells also induced fibronectin expression in the lung in a similar fashion to that of LLC cells (Supplementary Fig. 3d). Moreover, increased fibronectin expression was notable in multiple tissues exposed to MCM, such as the intestine and oviduct, consistent with the more aggressive metastatic nature of B16 cells (fibronectin expression: $P < 0.05$ days 3–5 and $P < 0.001$ days 7–9 (by ANOVA) in oviducts (Fig. 6a) and intestines (Fig. 6b) with MCM treatment compared with mice treated with LLC-conditioned media (LCM) or wild-type mice).

VEGFR1⁺ cells promote tumour adherence and growth

To confirm that VEGFR1⁺ progenitors promote the chemoattraction and attachment of circulating tumour cells, we isolated and red fluorescence-labelled (PKH26-Gl) VEGFR1⁺ cells from mice with malignancy (Supplementary Fig. 4). Within one hour of *in vitro* co-incubation with green fluorescence-labelled (PKH2-Gl) B16 or LLC cells, the HPCs aggregated, proliferated (150% increase) and promoted the attachment and proliferation of the tumour cells. In contrast, preculturing VEGFR1⁺ HPCs with either anti-VEGFR1 or anti-VLA-4 antibodies blocked this binding affinity and expansion (Supplementary Fig. 4a, middle and right panels). Using a transwell migration assay, tumour cells manifested enhanced mobility in response to bone marrow-derived VEGFR1⁺ cells (29.6 ± 1.4 tumour cells per $\times 200$ objective field) as compared to cells that do not express VEGFR1 (11.2 ± 0.4) and media alone (9.9 ± 0.9 , $P < 0.001$ by ANOVA; Supplementary Fig. 4b). The SDF-1/CXCR4 chemokine axis participates in homing and retention of HPCs within the bone marrow¹⁸. Specific tumour cell types, which express CXCR4, may also migrate in this fashion in response to local chemokine gradients^{19–21}. Within the fully formed pre-metastatic cluster containing VEGFR1⁺ cells, fibroblasts and fibronectin (Fig. 1a, left middle panel), SDF-1 (also known as CXCL12) became highly expressed (Supplementary Fig. 4c). We also observed CXCR4 expression in B16 melanoma and LLC tumours (Supplementary Fig. 4d). These data suggest that SDF-1 may provide one pathway for attracting CXCR4⁺ tumour cells to the pre-metastatic niche.

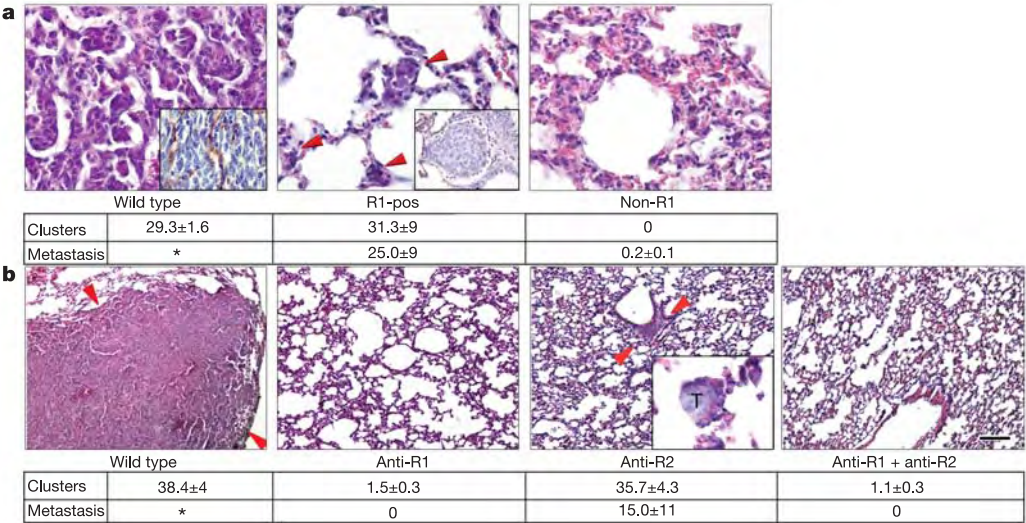


Figure 4 | Inhibition of homing of bone marrow cells prevents metastasis. **a**, VEGFR1⁺-selected bone marrow (R1-pos) permits micrometastasis (red arrows, middle panel) but prevents well-vascularized large metastases as seen in wild types (left panel), 24 days after LLC implantation. Insets show CD31 (endothelial marker) expression. Bone marrow depleted of VEGFR1⁺ cells (non-R1) abrogates both clusters and metastases (right panel) ($P < 0.01$ by ANOVA). The table shows the number of clusters and micrometastases per $\times 100$ objective field. *denotes that the metastasis filled the lung. (R1-pos, $n = 4$; non-R1, $n = 4$; wild type, $n = 6$; non-R1 plus wild

type, $n = 4$). **b**, Treatment with antibodies to VEGFR1 (anti-R1) and VEGFR2 (anti-R2) in mice with LLC tumours prevents both clusters and metastases ($P < 0.01$ by ANOVA; for all groups, $n = 5$). Arrows in the lung of the wild type denote a large LLC metastasis. Arrows in anti-R2 show a cluster, inset shows a micrometastasis within a cluster. T, tumour cells. The table shows the number of clusters and LLC micrometastases in lung per $\times 100$ objective field. *denotes that the metastasis filled the tissue. Scale bar at bottom right applies to panels **a** (20 μm ; wild type inset, 26 μm ; R1-pos inset, 32 μm) and **b** (40 μm ; anti-R2 inset, 20 μm).

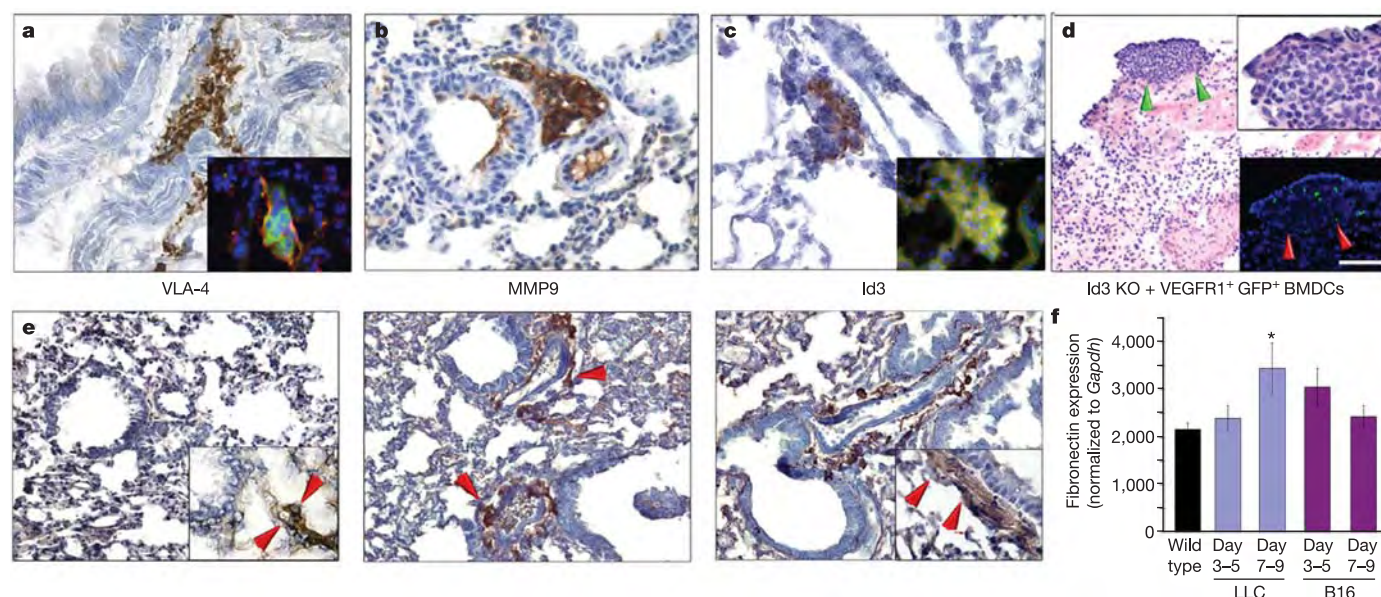


Figure 5 | The VLA-4/fibronectin pathway mediates cluster formation. **a–c**, Wild-type mice 14 days after tumour implantation develop clusters expressing VLA-4 (inset, VEGFR1 (red) and VLA-4 (green)), MMP9 and Id3 (inset, VEGFR1 (red) and Id3 (green)). **d**, Lung tissue in Id3 knockout (KO) mice with LLC tumours given VEGFR1⁺GFP⁺ BMDCs ($P < 0.01$ by ANOVA; $n = 6$). Green arrows show region in upper inset. Red arrows (lower inset) show the site of metastasis with GFP⁺VEGFR1⁺ cells. **e**, Baseline fibronectin expression in the wild-type lung ($n = 6$) (left panel). Increased stromal fibronectin in the peribronchial region of the

pre-metastatic lung at day three (middle panel, arrows), with maximal expression on day 14 (right panel). Insets, PDGFR α expression indicates resident fibroblasts laying down fibronectin. **f**, Quantitative RT-PCR reveals increased fibronectin expression in the lungs of mice with LLC tumours compared with wild type ($*P < 0.05$ by ANOVA; $n = 6$), and a similar earlier trend in lungs from animals with B16 melanoma. Scale bar at top right applies to panels **a**, **b**, **c** (40 μ m; insets, 8 μ m), **d** (80 μ m; top right inset, 20 μ m; bottom right inset 80 μ m) and **e** (40 μ m; insets, 20 μ m).

Tumour-derived conditioned media dictate metastatic patterns

To delineate the mechanism of the organ-specific metastatic potential of LLC and B16 cells, we collected culture-derived conditioned media. Similarly, but more rapidly than primary LLC cells, the intraperitoneal injection of LCM generated fibronectin expression,

possibly from resident fibroblasts, and BMDC cluster formation (Supplementary Fig. 5a) compared with media alone (Supplementary Fig. 5a, insets). MCM stimulated fibronectin expression to a greater extent in liver than LCM (Supplementary Fig. 5b). MCM caused enhanced fibroblast proliferation (data not shown) and

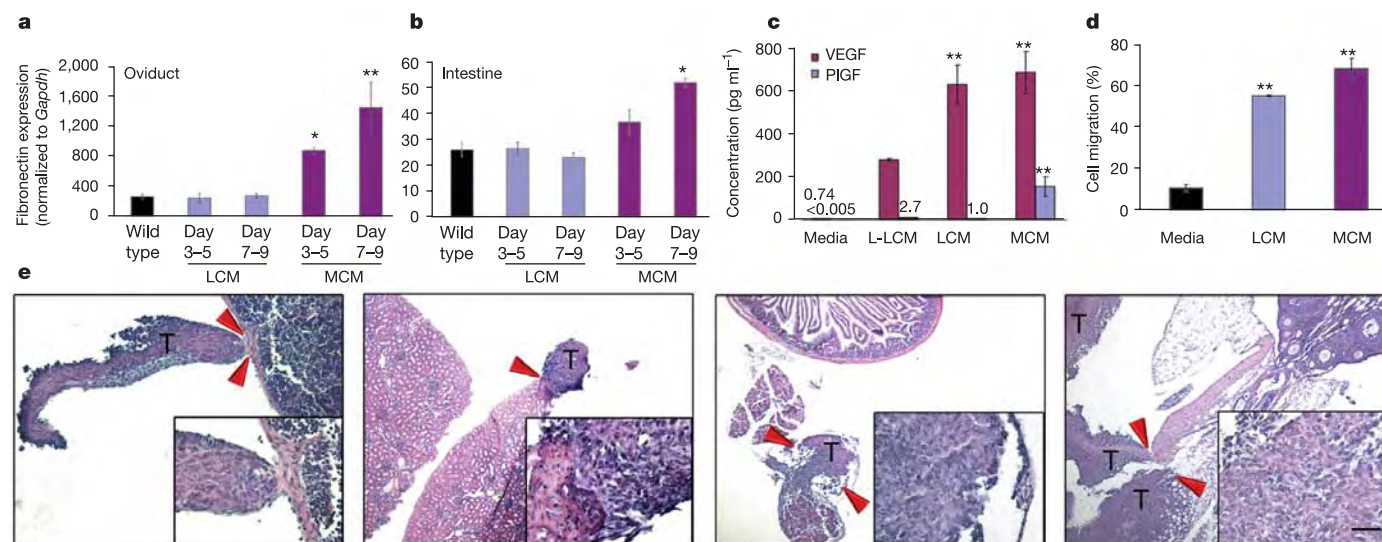


Figure 6 | Redirection of LLC metastases to atypical sites. **a, b**, By quantitative RT-PCR analysis, increased fibronectin expression was seen in the oviduct (**a**) and intestine (**b**) in mice given MCM compared with wild-type and LCM-treatment. For oviduct, $*P < 0.05$ at days 3–5 and $**P < 0.001$ for days 7–9 compared with wild type, and for intestine, $*P < 0.001$ at days 7–9 compared with wild type by ANOVA ($n = 6$). **c**, ELISA assay (in triplicate) for VEGF and PlGF levels in the conditioned media ($*P < 0.05$ when compared with L-LCM, $**P < 0.01$ when compared

with media alone, by ANOVA). **d**, Transwell migration assays (in triplicate) demonstrate enhanced migration of VEGFR1⁺ cells to LCM and MCM ($**P < 0.001$ by ANOVA). **e**, Treatment with MCM redirects the metastatic spread of LLC to B16 melanoma metastatic sites, such as the spleen (left panel), kidney (left middle panel), intestine (right middle panel) and oviduct (right panel). Arrows denote the regions of metastatic borders, which are shown in the insets ($n = 6$). T, LLC tumour cells. Scale bar at bottom right applies to panel **e** (200 μ m; insets, 20 μ m).

fibronectin expression with cluster formation in a wide range of organs, as shown for intestine (Fig. 6a, b; Supplementary Fig. 5b) in comparison to media (Supplementary Fig. 5b, inset). We analysed LCM and MCM for variations in growth factors to account for the distinct metastatic potentials and profiles of LLC and B16 (Fig. 6c). We found high levels of VEGF in both conditioned media, more than in plasma from tumour-bearing mice (Supplementary Fig. 5c). However, in MCM and melanoma-derived plasma we specifically detected higher levels of placental growth factor (PlGF), which signals through VEGFR1 alone, as compared with LCM- and LLC-derived plasma (Fig. 6c, Supplementary Fig. 5c). Furthermore, in the low-metastatic-variant of LLC, levels of both VEGF and PlGF were much lower in the conditioned media (L-LCM) and plasma compared with its more aggressive counterpart (Fig. 6c, Supplementary Fig. 5c). In a transwell assay, LCM and MCM enhanced the migration of VEGFR1⁺ BMDCs most effectively when compared with the other growth factor conditions (LCM 55% $\pm$ 0.4, MCM 68.1% $\pm$ 5, media 10.8% $\pm$ 1.7, $P < 0.001$ by ANOVA; Fig. 6d). Considering these results, we questioned whether cytokines such as PlGF present in MCM were capable of redirecting LLC metastases to non-conventional metastatic sites for this tumour. MCM given before intradermal LLC implantation, and daily thereafter, resulted in the redirection of LLC metastasis from lung to those sites frequently observed in melanoma including kidney, spleen, intestine and oviduct (Fig. 6e). Our results demonstrate that tumour-specific chemokines and/or cytokines present in conditioned media, along with the VEGFR1⁺ cellular clusters, are another determinant in the multidimensional programme driving metastatic spread.

The precise cellular and molecular mechanisms that dictate metastasis of a specific tumour to a predetermined metastatic location are not known. Many tumours have a predilection for metastasis to specific organs. Based on the current dogma, metastatic predisposition is believed to reflect inherent molecular differences in tumour cells themselves and the potential influence by surrounding stromal cells, which include the vasculature, connective tissue and immune cells^{22–26}. Our results introduce the concept that tumour metastasis is initiated by a well-defined sequence of events dependent on cellular ‘bookmarking’ through site-specific delivery of VEGFR1⁺ cells to form permissive niches within target organs. Our data suggest that differences in tumour-secreted humoral factors promote metastatic spread in specific distant organs. Within days following tumour implantation, fibronectin becomes upregulated in certain locations by resident fibroblast and fibroblast-like cells within target organs that are conventional sites of metastasis, corresponding to the particular primary tumour. Simultaneously, HPCs exit the bone marrow into the peripheral circulation as previously described¹¹. As a result of the niche-specific directional cues from fibronectin, VEGFR1⁺ HPCs, expressing VLA-4 and Id3, can traverse established endothelium to form a pre-metastatic niche before the arrival of CXCR4⁺ tumour cells and VEGFR2⁺ endothelial cells. These clusters, with MMP9 production altering the microenvironment and enhanced expression of SDF-1 creating a chemokine gradient, permit the attraction of tumour cells and their incorporation into the niche, thereby developing a complete metastatic lesion. We show that inhibition by a VEGFR1 antibody or depletion of VEGFR1⁺ cells from the bone marrow prevents the formation of pre-metastatic clusters and, therefore, metastases. Moreover, blocking either VEGFR1 or VLA-4 inhibits the binding and establishment of the haematopoietic cell clusters and tumour cells. Restoration of the pre-metastatic niche and metastasis with the introduction of wild-type VEGFR1⁺ cells into Id3 knockout mice suggests that the expression of Id3 induces expression of the necessary elements, including MMP9, integrins and possibly chemokines, to provide a road map for the homing of VEGFR1⁺ cells essential for the establishment of the pre-metastatic niche.

Much focus has been placed on the role of inflammatory cells in aiding in tumour adherence and invasion into distant organs^{27–30}.

The VEGFR1⁺ HPCs identified in this study show characteristics common to physiological pathways of inflammation by providing the necessary adhesion molecules, proteinases, chemokines and growth conditions to create a conducive microenvironment for engraftment of tumour cells^{12,20,31}. The pre-metastatic niche, however, is distinct, introducing an undifferentiated state as seen with the VEGFR1⁺ HPC population. This is the first direct evidence that a non-neoplastic cell population can portend a future metastatic site. Furthermore, identification of haematopoietic clusters in human tissues before evidence of tumour spread demonstrates the applicability of targeting VEGFR1 and VLA-4 to identify and prevent metastasis in the clinical setting. This concept will have a tremendous impact on tumour staging, and may alter the landscape of adjuvant chemotherapy.

METHODS

Bone marrow transplantation. Wild-type C57Bl/6 mice were lethally irradiated (950 rads) and transplanted with 1×10^6 β -gal⁺ bone marrow cells (from Rosa26 mice) or 1×10^6 GFP⁺ bone marrow cells (from EGFP-transgenic mice, C57Bl/6-TgN(ActbEGFP)10sb/J; Jackson Laboratory)². After 4 weeks, mice were injected intradermally in the flank with either 2×10^6 LLC or B16 cells (American Type Culture Collection).

Selective bone marrow transplantation. Mice irradiated as described above received a bone marrow transplant from purified cell populations obtained as described in the Supplementary Methods.

β -Galactosidase staining. Tissues and femoral bones were fixed in 4% paraformaldehyde for 4 h. The samples were stained in 5-bromo-4-chloro-3-indolyl- β -D-galactoside (X-gal) solution at 37°C, as described³², for 36 h and then embedded².

GFP visualization. Tissues were immediately frozen in OCT compound (Tissue-Tek) without fixation. Serial sections (cryostat, Leica) were mounted with Vectashield containing DAPI (4,6-diamidino-2-phenylindole), and visualized with an ultraviolet fluorescent microscope (Nikon Eclipse E800) with a Retiga camera (QImaging) through IPLab version 3.65a imaging software (Scanalytics).

Immunohistochemistry. Tissues were fixed and embedded in OCT or paraffin as previously described¹⁶. The following antibodies were used: VEGFR1 clone MF-1 (ImClone Systems) or Flt1 clone C-17 (Santa Cruz Biotechnology); CD31 SC-1506 (Santa Cruz Biotechnology); VEGFR2 DC101 (ImClone Systems); MMP9 D19557 (Oncogene); Id3 C-20 (Santa Cruz Biotechnology); Fibronectin TV-1 (Chemicon); CD11b CBRM1/5 (eBioscience); CD34 RAM34 (BD Pharmingen); c-Kit ACK2 (eBioscience); PDGFR α APA5 (BD Pharmingen); α V (Chemicon); CD133 13A4 (eBioscience); α 4/VLA-4 PS-2 (Southern Biotech); α 5 (CD49e, 5H10-27); α 6/CD49f GoH3 (BD Pharmingen); β 1 9EG7 (BD Pharmingen); β 2 M18/2 (BD Pharmingen); β 4 (Santa Cruz Biotechnology); β 7 M293 (BD Pharmingen); SDF-1 79018.111 (R&D Systems); and CXCR4 2B11 (BD Pharmingen).

Double immunofluorescence. Tissues in OCT were post-fixed with acetone. A double immunofluorescence protocol was performed as described in the Supplementary Methods.

Antibody targeting. Wild-type mice were inoculated with 2×10^6 LLC or B16 cells. For blockade of VEGFR1 function, mice were injected intraperitoneally every 48 h, between day 7–22, with rat anti-mouse VEGFR1 antibody (MF-1, IgG1, 400 μ g, ImClone Systems) or VEGFR2 antibody (DC101, IgG1, 800 μ g, ImClone Systems) or both, or with IgG control antibody, and then killed on day 24.

Conditioned media assays. Conditioned media was filtered (0.22- μ m filter) from serum-free media cultured on B16 (MCM) or LLC (LCM) cells for 18 h, as described³³. Conditioned media (300 μ l) was injected intraperitoneally daily for nine days into wild-type mice that had received Rosa26 bone marrow transplants four weeks earlier. Tissues were stained for fibronectin TV-1 (Chemicon) and β -gal. For tumour redirection studies, intraperitoneal injections of MCM (300 μ l) commenced two days before intradermal LLC implantation and then daily over the next 21 days. Matched control groups with and without tumour were given serum-free media. Wild-type mice were injected with MCM (300 μ l) daily for seven days before tail vein injection of B16 tumour cells, and then daily until killed either one or four days after intravenous tumour administration. Lungs were perfused with PBS before embedding in OCT.

Migration assays. Migration of VEGFR1⁺ cells in response to conditioned media was measured in a transwell assay. VEGFR1⁺ cells were isolated as above, and 1×10^5 cells suspended in serum-free media placed in the upper compartment of 5- μ m-pore transwells (Costar, Corning). Cells were allowed to migrate for 18 h with conditioned media or corresponding control media in the lower

compartment, with the analysis of cell counts assessed every 6 h using a haemocytometer and trypan blue.

Quantitative analysis of fibronectin expression. Lung tissue was homogenized with a tissue homogenizer in TriZol reagent, and RNA was extracted as described previously³⁴. Fibronectin gene expression was quantified and normalized to glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*) expression by polymerase chain reaction with reverse transcription (RT-PCR) using TaqMan gene expression assays (Applied Biosystems) as described previously³⁵.

Chemokine assays. Conditioned media, serum-free media and plasma obtained from mice with day 14 tumours were analysed for VEGF and PlGF concentrations by an enzyme-linked immunosorbent assay (ELISA; Quantikine, R&D Systems) according to the manufacturer's instructions.

Flow cytometry. Flow cytometry was performed on an entire right lung after perfusion with PBS by right-ventricular injection. The tissue was minced into small pieces, filtered with 100- and 40- μ m filters (BD Biosciences) to form a single-cell suspension as previously described^{35,36}.

Human specimens. Human specimens include: tumour tissue, adjacent normal tissue (beyond tumour margins), distant normal tissue and lymph nodes. Tissues were embedded as described above and stained with antibodies to human VEGFR1 FB5 (ImClone Systems) or Flt1 (Calbiochem). Tissue samples were obtained and handled in accordance with an approved Institutional Review Board application.

Quantitative immunohistochemistry. Using both IPLab and Adobe Photoshop 7.0, random $\times 100$ objective fields were analysed by selecting a standardized colour range for β -gal or immunohistochemical staining. After boundary delineation, the area under the pixelation histogram was calculated, comparing total staining area to total tissue area.

Statistical analyses. Results are expressed as mean $\pm$ s.e.m. Data were analysed by Student's *t*-test and one way analysis of variance (ANOVA) using the GraphPad Prism statistical program. *P* values < 0.05 were considered significant. Error bars depict s.e.m.

Received 13 May; accepted 19 August 2005.

- Coussens, L., Tinkle, C., Hanahan, D. & Werb, Z. MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis. *Cell* **103**, 481–490 (2000).
- Lyden, D. *et al.* Impaired recruitment of bone-marrow-derived endothelial and hematopoietic precursor cells blocks tumour angiogenesis and growth. *Nature Med.* **7**, 1194–1201 (2001).
- Autiero, M., Luttun, A., Tjwa, M. & Carmeliet, P. Placental growth factor and its receptor, vascular endothelial growth factor receptor-1: novel targets for stimulation of ischemic tissue revascularization and inhibition of angiogenic and inflammatory disorders. *J. Thromb. Haemost.* **1**, 1356–1370 (2003).
- Neeson, P., Thurlow, P., Jamieson, G. & Bradley, C. Lymphocyte-facilitated tumour cell adhesion to endothelial cells: the role of high affinity leukocyte integrins. *Pathology* **35**, 50–55 (2003).
- Hattori, K. *et al.* Placental growth factor reconstitutes hematopoiesis by recruiting VEGFR1⁺ stem cells from bone-marrow microenvironment. *Nature Med.* **8**, 841–849 (2002).
- Pollard, J. W. Tumour-educated macrophages promote tumour progression and metastasis. *Nature Rev. Cancer* **4**, 71–78 (2004).
- Hiratsuka, S. *et al.* MMP9 induction by vascular endothelial growth factor receptor-1 is involved in lung-specific metastasis. *Cancer Cell* **2**, 289–300 (2002).
- De Palma, M., Vinneri, M. A., Roca, C. & Naldini, L. Targeting exogenous genes to tumour angiogenesis by transplantation of genetically modified hematopoietic cells. *Nature Med.* **9**, 789–795 (2003).
- Burger, J., Spoo, A., Dwenger, A., Burger, M. & Behringer, D. CXCR4 chemokine receptors (CD184) and $\alpha_4\beta_1$ integrins mediate spontaneous migration of human CD34⁺ progenitors and acute myeloid leukaemia cells beneath marrow stromal cells (pseudoemperipolesis). *Br. J. Haematol.* **122**, 579–589 (2003).
- Scott, L., Priestly, G. & Papayannopoulou, T. Deletion of α_4 integrins from adult hematopoietic cells reveals roles in homeostasis, regeneration, and homing. *Mol. Cell. Biol.* **23**, 9349–9360 (2003).
- Jonjic, N. *et al.* Molecules involved in the adhesion and cytotoxicity of activated monocytes on endothelial cells. *J. Immunol.* **148**, 2080–2083 (1992).
- Heissig, B. *et al.* Recruitment of stem and progenitor cells from the bone marrow niche requires MMP-9 mediated release of kit-ligand. *Cell* **109**, 625–637 (2002).
- Bergers, G. *et al.* Matrix Metalloproteinase-9 triggers the angiogenic switch during carcinogenesis. *Nature Cell Biol.* **2**, 737–744 (2000).
- Huhtala, P. *et al.* Cooperative signalling by $\alpha_5\beta_1$ and $\alpha_4\beta_1$ integrins regulates metalloproteinase gene expression in fibroblasts adhering to fibronectin. *J. Cell Biol.* **129**, 867–879 (1995).
- Yakubenko, V. P., Lobb, R. R., Plow, E. F. & Ugarova, T. P. Differential induction of gelatinase B (MMP-9) and gelatinase A (MMP-2) in T lymphocytes upon $\alpha_4\beta_1$ -mediated adhesion to VCAM-1 and the CS-1 peptide of fibronectin. *Exp. Cell Res.* **260**, 73–84 (2000).
- Ruzinova, M. B. *et al.* Effect of angiogenesis inhibition by Id loss and the contribution of bone-marrow-derived endothelial cells in spontaneous murine tumours. *Cancer Cell* **4**, 277–289 (2003).
- Lyden, D. *et al.* Id1 and Id3 are required for neurogenesis, angiogenesis and vascularization of tumour xenografts. *Nature* **401**, 670–677 (1999).
- Ratajczak, M. Z. *et al.* Stem cell plasticity revisited: CXCR4-positive cells expressing mRNA for early muscle, liver and neural cells 'hide out' in the bone marrow. *Leukemia* **18**, 29–40 (2004).
- Lapidot, T. & Petit, I. Current understanding of stem cell mobilization: The roles of chemokines, proteolytic enzymes, adhesion molecules, cytokines and stromal cells. *Exp. Hematol.* **30**, 973–981 (2002).
- Balkwill, F. The significance of cancer cell expression of the chemokine receptor CXCR4. *Semin. Cancer Biol.* **14**, 171–179 (2004).
- Muller, A. *et al.* Involvement of chemokine receptors in breast cancer metastasis. *Nature* **410**, 50–56 (2001).
- Hynes, R. O. Metastatic potential: generic predisposition of the primary tumour or rare, metastatic variants—or both? *Cell* **113**, 821–823 (2003).
- Bergers, G., Song, S., Meyer-Morse, N., Bergsland, E. & Hanahan, D. Benefits of targeting both pericytes and endothelial cells in the tumour vasculature with kinase inhibitors. *J. Clin. Invest.* **111**, 1287–1295 (2003).
- Fidler, I. The organ microenvironment and cancer metastasis. *Differentiation* **70**, 498–505 (2002).
- Duda, D. G. *et al.* Differential transplantability of tumour-associated stromal cells. *Cancer Res.* **64**, 5920–5924 (2004).
- Folkman, J. Role of angiogenesis in tumour growth and metastasis. *Semin. Oncol.* **29**, 515–518 (2002).
- Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).
- Borsig, L., Wong, R., Hynes, R. O., Varki, N. M. & Varki, A. Synergistic effects of L- and P-selectin in facilitating tumour metastasis can involve non-mucin ligands and implicate leukocytes as enhancers of metastasis. *Proc. Natl Acad. Sci. USA* **99**, 2193–2198 (2000).
- Lin, E. Y., Ngyuen, A. V., Russell, R. G. & Pollard, J. W. Colony stimulating factor 1 promoted progression of mammary tumours to malignancy. *J. Exp. Med.* **193**, 727–740 (2001).
- Qian, F., Hanahan, D. & Weissman, I. L. L-selectin can facilitate metastasis to lymph nodes in a transgenic mouse model of carcinogenesis. *Proc. Natl Acad. Sci. USA* **98**, 3976–3981 (2002).
- Schoppmann, S. *et al.* Tumour-associated macrophages express lymphatic endothelial growth factors and are related to peritumoural lymphangiogenesis. *Am. J. Pathol.* **161**, 947–956 (2002).
- Tam, P. P., Parameswaran, M., Kinder, S. J. & Weinberger, R. P. The allocation of epiblast cells to the embryonic heart and other mesodermal lineages: the role of ingression and tissue movement during gastrulation. *Development* **124**, 1631–1642 (1999).
- Kessinger, A., Mann, S., Murphy, B. O., Jackson, J. D. & Sharp, J. G. Circulating factors may be responsible for murine strain-specific responses to mobilizing cytokines. *Exp. Hematol.* **29**, 775–778 (2001).
- Hashimoto, N., Jin, H., Liu, T., Chensue, S. W. & Phan, S. H. Bone marrow-derived progenitor cells in pulmonary fibrosis. *J. Clin. Invest.* **113**, 243–252 (2004).
- Jensen, K. K. *et al.* The human herpes virus 8-encoded chemokine receptor is required for angioproliferation in a murine model of Kaposi's sarcoma. *J. Immunol.* **174**, 3686–3694 (2005).
- Huax, F., Tianju, L., McGarry, B., Ullenbruch, M. & Phan, S. H. Dual roles of IL-4 in lung injury and fibrosis. *J. Immunol.* **170**, 2083–2092.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Barna for critical reading of the manuscript and L. Breda, S. Rivella and S. Neustein for discussions. R.N.K. is a recipient of the Laura Rosenberg Fellowship award and supported by a grant from the American Hellenic Educational Progressive Association (Fifth District) and the LTC Foundation. D.L. is supported by the Doris Duke Charitable Foundation, the Children's Blood Foundation, the Emerald Foundation, the Theodore A. Rapp Foundation and a grant from the National Cancer Institute. S.R. is an investigator of the Howard Hughes Medical Institute and supported by grants from the American Cancer Society, the Leukemia and Lymphoma Society, and the National Institutes of Health.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.L. (dcl2001@med.cornell.edu) or S.R. (srafi@med.cornell.edu).

LETTERS

Measurement-induced entanglement for excitation stored in remote atomic ensembles

C. W. Chou¹, H. de Riedmatten¹, D. Felinto¹, S. V. Polyakov¹, S. J. van Enk² & H. J. Kimble¹

A critical requirement for diverse applications in quantum information science is the capability to disseminate quantum resources over complex quantum networks^{1,2}. For example, the coherent distribution of entangled quantum states together with quantum memory (for storing the states) can enable scalable architectures for quantum computation³, communication⁴ and metrology⁵. Here we report observations of entanglement between two atomic ensembles located in distinct, spatially separated set-ups. Quantum interference in the detection of a photon emitted by one of the samples projects the otherwise independent ensembles into an entangled state with one joint excitation stored remotely in 10^5 atoms at each site⁶. After a programmable delay, we confirm entanglement by mapping the state of the atoms to optical fields and measuring mutual coherences and photon statistics for these fields. We thereby determine a quantitative lower bound for the entanglement of the joint state of the ensembles. Our observations represent significant progress in the ability to distribute and store entangled quantum states.

Entanglement is a uniquely quantum mechanical property of the correlations among various components of a physical system. Initial demonstrations of entanglement were made for photon pairs from the fluorescence in atomic cascades^{7,8} and from parametric down-conversion⁹. More recently, entanglement has been recognized as a critical resource for accomplishing tasks that are otherwise impossible in the classical domain¹. Spectacular advances have been made in the generation of quantum entanglement for diverse physical systems^{1,2}, including entanglement stored for many seconds in trapped ions for distances on the millimetre scale^{10,11}, long-lived entanglement of macroscopic quantum spins persisting for milliseconds on the centimetre scale¹², and remote entanglement carried by photon pairs over distances of tens of kilometres of optical fibres¹³.

For applications in quantum information science, entanglement can be created deterministically by precisely controlling quantum dynamics for a physical system, or probabilistically by quantum interference in a suitable measurement with random instances of success. In the latter case, it is essential that success be heralded unambiguously so that the resulting entangled state is available for subsequent use. In either case, quantum memory is required to store the entangled states until they are required for the protocol at hand.

There are by now several examples of entanglement generated 'on demand'¹, beginning with the realization of the Einstein–Podolsky–Rosen (EPR) paradox for continuous quantum variables¹⁴ and the deterministic entanglement of the discrete internal states of two trapped ions¹⁵. Important progress has been made towards measurement-induced entanglement on various fronts, including the observation of entanglement between a trapped ion and a photon (ref. 16 and references therein).

Here, we report the initial observation of entanglement created probabilistically from quantum interference in the measurement process, with the resulting entangled state heralded unambiguously and stored in quantum memory for subsequent use. As illustrated in Fig. 1, the detection of a photon from either of two atomic ensembles (L, R) in an indistinguishable fashion results in an entangled state with one 'spin' excitation shared at a distance of 2.8 m between the ensembles and distributed symmetrically among $\sim 10^5$ atoms at each site⁶. Confirmation of entanglement is achieved by mapping this stored excitation onto light fields after 1- μ s delay^{6,17} and by suitable measurements of the quantum statistics of the resulting optical fields. Our results provide the first realization of the capability to transfer a stored entangled state of matter to an entangled state of light.

Our experiment is motivated by the protocol of Duan, Lukin, Cirac and Zoller (DLCZ)⁶ for the realization of scalable quantum communication networks with atomic ensembles. The DLCZ protocol introduced a number of ideas for quantum information processing and is the subject of active investigation. In this direction, nonclassical correlations^{17–24} and entanglement²⁵ have been observed between pairs of photons emitted by a single atomic ensemble. Observations of coherence between two cylindrical volumes of cold rubidium atoms within a single magneto-optical trap have also been reported²⁶, although entanglement was not demonstrated between the two regions^{27,28}.

A simple schematic of our experiment is given in Fig. 1, with further details provided in refs 17, 21 and 23. For the writing stage of the protocol, two classical pulses traverse the L and R ensembles in parallel and generate fields 1_L , 1_R by spontaneous Raman scattering (see Fig. 1a). The intensity of the pulses is made sufficiently weak that the probability of creating more than one excitation in the symmetric collective mode⁶ of the ensemble is very low²¹.

Entanglement between the L and R ensembles is created by combining the output fields 1_L , 1_R on the beamsplitter BS₁, with outputs directed to two photodetectors D_{1a}, D_{1b} (see Fig. 1a). For small excitation probability and with unit overlap of the fields at BS₁, a detection event at D_{1a} or D_{1b} arises indistinguishably from either field 1_L or 1_R , so that the L and R ensembles are projected into an entangled state, which in the ideal case can be written as^{6,29}:

$$|\Psi_{L,R}\rangle = \epsilon_L |1\rangle_L |0\rangle_R \pm e^{i\eta_1} \epsilon_R |0\rangle_L |1\rangle_R \quad (1)$$

where $|0\rangle_{L,R}$, $|1\rangle_{L,R}$ refers to the two ensembles L and R with 0 and 1 collective excitations respectively, ϵ_L (or ϵ_R) is the normalized amplitude of photon generation from ensemble L (or R), and the sign (+ or –) is set by whichever detector records the event. The phase $\eta_1 = \Delta\beta_w + \Delta\gamma_1$, where $\Delta\beta_w$ is the phase difference of the write beams at the L and R ensembles, and $\Delta\gamma_1$ is the phase

¹Norman Bridge Laboratory of Physics 12-33, California Institute of Technology, Pasadena, California 91125, USA. ²Bell Labs, Lucent Technologies, Room 1D-428, 600-700 Mountain Avenue, Murray Hill, New Jersey 07974, USA.

difference acquired by the 1_L and 1_R fields in propagation from the ensembles to the beamsplitter BS_1 . We note that to achieve entanglement as in equation (1), η_1 has to be kept constant from trial to trial.

To verify the entanglement, we map the delocalized atomic excitation into a field state by applying simultaneously strong read beams at the two ensembles (see Fig. 1b). If the state transfer were to succeed with unit probability, the conditional state $|\Psi_{L,R}\rangle$ of the ensembles would be mapped to an entangled state of two modes for the Stokes fields 2_L and 2_R given in the ideal case by^{6,29}:

$$|\Phi_{LR}\rangle = \epsilon_L |1\rangle_{2L} |0\rangle_{2R} \pm e^{i(\eta_1 + \eta_2)} \epsilon_R |0\rangle_{2L} |1\rangle_{2R} \quad (2)$$

where $|0\rangle_{2L,2R}$, $|1\rangle_{2L,2R}$ refer to the Raman fields 2_L , 2_R with 0, 1 photons, respectively. Here, $\eta_2 = \Delta\beta_r + \Delta\gamma_2$, where $\Delta\beta_r$ is the phase difference of the read beams at the L and R ensembles, and $\Delta\gamma_2$ is the phase difference acquired by the 2_L and 2_R fields in propagation from the ensembles to the beamsplitter BS_2 in Fig. 1b. In our experiment, the phases η_1 and η_2 can be independently controlled and are actively stabilized by utilizing auxiliary fields at $1.06\ \mu\text{m}$ that co-propagate

along the paths of the write and read beams and of the 1_L , 1_R and 2_L , 2_R fields.

Of course, the states in equations (1) and (2) are idealizations that must be generalized to describe our actual experiment^{6,27,29}. Specifically, the presence of various sources of noise necessarily transforms these pure states into mixed states. Equations (1) and (2) also neglect the vacuum contribution as well as higher-order terms, which are intrinsic to DLCZ protocol and which otherwise arise from diverse experimental imperfections. Moreover, the above analysis assumes that all excitations are in the correct 'modes' (both for optical fields and for the collective atomic 'spin flips'), that excitations of the ensembles map one-to-one to photons in fields 1 and 2, and that diverse sources of background light are absent.

The procedure that we have devised to provide a robust, model-independent determination of entanglement is based upon quantum tomography of the 2_L and 2_R fields (see Supplementary Information for details). Because entanglement cannot be increased by local operations on either of the two ensembles, the entanglement for the state of the ensembles will be always greater than or equal to that

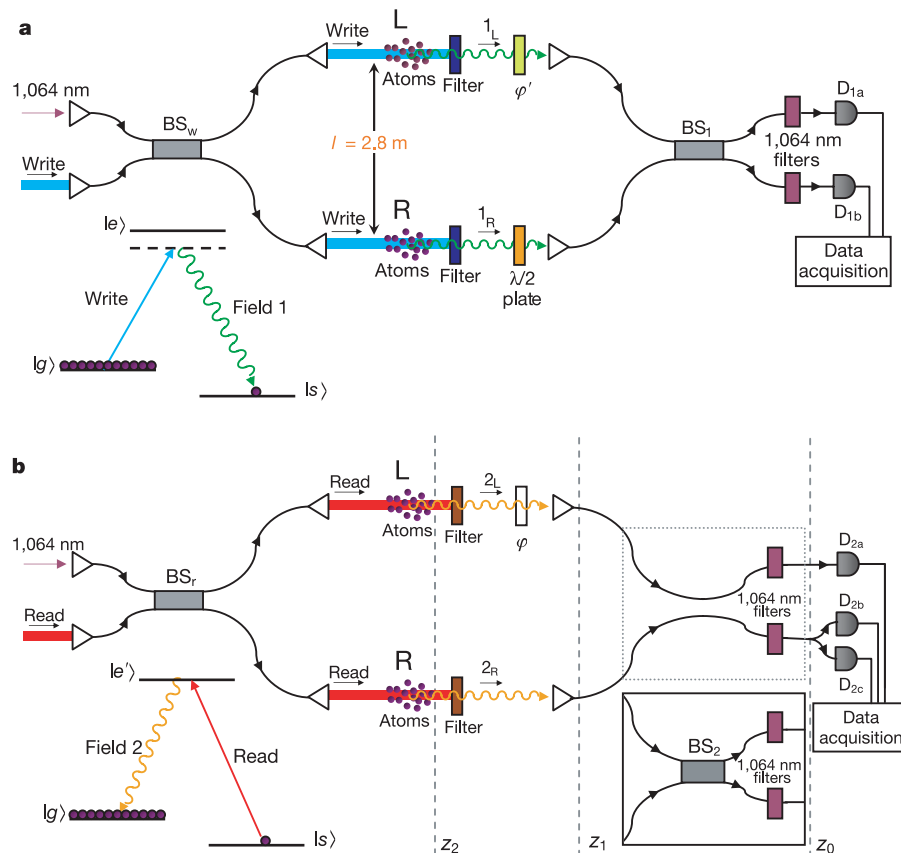


Figure 1 | An overview of our experiment to entangle two atomic ensembles is shown. a, Set-up for generating entanglement between two pencil-shaped ensembles L and R located within spherical clouds of cold caesium atoms. The atomic level structure for the writing process consists of the initial ground state $|g\rangle$ ($6S_{1/2}$, $F = 4$ level of atomic caesium), the ground state $|s\rangle$ for storing a collective spin flip ($6S_{1/2}$, $F = 3$ level), and the excited level $|e\rangle$ ($6P_{3/2}$, $F' = 4$). The transition $|g\rangle \rightarrow |e\rangle$ in each ensemble is initially coupled by a write pulse detuned from resonance to generate the forward-scattered anti-Stokes field 1 from the transition $|e\rangle \rightarrow |s\rangle$. The L and R ensembles are excited by synchronized writing pulses obtained from beamsplitter BS_w . After filtering, the anti-Stokes fields 1_L and 1_R are collected, coupled to fibre-optic channels, and interfere at beamsplitter BS_1 , with outputs directed towards two single-photon detectors D_{1a} and D_{1b} . **b**, Schematic for verification of entanglement between the L and R ensembles by conversion of atomic to field excitation by way of simultaneous read pulses obtained

from BS_r . The read pulses reach the samples after a programmable delay from the write pulses, and couple the transition $|s\rangle \rightarrow |e\rangle$ ($|e\rangle$ being the $6P_{1/2}$, $F' = 4$ level), leading to the emission of the forward-scattered Stokes fields 2_L and 2_R from the transition $|e\rangle \rightarrow |g\rangle$. The upper inset shows the configuration used to measure the diagonal elements p_{ij} of $\hat{\rho}_{2L,2R}$ in equation (3) from the photo-detection events at D_{2a} , D_{2b} and D_{2c} . Reconfiguring the fibre connections, we can easily pass from the configuration of the upper inset to the one of the lower inset, which is used to generate interference of the 2_L and 2_R fields at beamsplitter BS_2 to measure the off-diagonal coherence d in $\hat{\rho}_{2L,2R}$. In **a** and **b**, the incident write and read beams are orthogonally polarized and combined at a polarizing beamsplitter (not shown), and are focused to a waist of about $30\ \mu\text{m}$ in the sample region. All beamsplitters BS are polarization-maintaining fibre beamsplitters. The $\sim 12\ \text{m}$ arms of both write and read interferometers are actively stabilized using an auxiliary Nd:YAG laser at $1.06\ \mu\text{m}$.

measured for the state of the light fields. Specifically, conditioned upon a detection at D_{1a} or D_{1b} , we consider the density matrix:

$$\tilde{\rho}_{2L,2R} = \frac{1}{\tilde{P}} \begin{pmatrix} p_{00} & 0 & 0 & 0 \\ 0 & p_{01} & d & 0 \\ 0 & d^* & p_{10} & 0 \\ 0 & 0 & 0 & p_{11} \end{pmatrix} \quad (3)$$

which is written in the basis $|n\rangle_{2L}|m\rangle_{2R}$, with the number of photons $\{n, m\} = \{0, 1\}$. p_{ij} is then the probability to find i photons in mode 2_L and j photons in mode 2_R , and d gives the coherence between the $|1\rangle_{2L}|0\rangle_{2R}$ and $|0\rangle_{2L}|1\rangle_{2R}$ states. $\tilde{\rho}_{2L,2R}$ is obtained from the full density matrix $\rho_{2L,2R}$ by restricting it to the subspace where there is at most one photon in each mode, with then $\tilde{P} = p_{00} + p_{01} + p_{10} + p_{11}$. The concurrence $C(\tilde{\rho}_{2L,2R})$ for $\tilde{\rho}_{2L,2R}$ provides a lower bound for the concurrence $C(\rho_{2L,2R})$ for $\rho_{2L,2R}$ [$C(\rho_{2L,2R}) \geq \tilde{P}C(\tilde{\rho}_{2L,2R})$], so we devise measurements to deduce the various components of $\tilde{\rho}_{2L,2R}$. The concurrence $C(\tilde{\rho}_{2L,2R})$ can then be calculated from equation (3) by way of ref. 30:

$$\tilde{P}C = \max(2|d| - 2\sqrt{p_{00}p_{11}}, 0) \quad (4)$$

The entanglement of formation E follows directly from C , where E and C both range from 0 to 1 for our system and E is a monotonically increasing function of C (ref. 30).

As a first step in the determination of C we measure the diagonal elements p_{ij} . As shown in Fig. 1b, the field-2 output of each ensemble is directed to different sets of detectors in order to record photon-counting probabilities for the fields $2_L, 2_R$ separately. From the record of photoelectric counting events, we then deduce the diagonal elements of $\tilde{\rho}_{2L,2R}$, which are listed in Table 1. From equation (4) and noting that $|d|^2 \leq p_{10}p_{01}$, a necessary requirement for $C > 0$ is that there be a suppression of two-photon events relative to the square of the probability for single-photon events for the fields $2_L, 2_R$, that is: $h_c^{(2)} \equiv p_{11}/(p_{10}p_{01}) < 1$. For our measurements, we find $h_c^{(2)} = 0.30 \pm 0.04$ for events conditioned on detection at D_{1a} , and $h_c^{(2)} = 0.35 \pm 0.04$ for events conditioned on D_{1b} (ref. 21). In contrast, for non-conditioned events, we find $h_{nc}^{(2)} = 0.99 \pm 0.04$.

The second step in our tomography protocol is to determine the coherence term d in equation (3), which we accomplish by adding a relative phase shift φ for the fields $2_L, 2_R$, and by combining them at the beamsplitter BS_2 shown in Fig. 1b. By recording the conditional count rate after the beam splitter as function of φ , we can measure an interference fringe with a visibility V , with $|d|$ then following from V and the p_{ij} . Roughly, for 50/50 beamsplitters and neglecting higher-order terms (that are employed in our actual analysis), we would have $|d| \cong V(p_{10} + p_{01})/2$.

Figure 2 shows conditional counts $N_{2a}, N_{2b} + N_{2c}$ as functions of φ . These data demonstrate that the indistinguishable character of measurement events at detectors D_{1a} (Fig. 2a) and D_{1b} (Fig. 2b) induces a high degree of coherence between the otherwise independent ensembles L, R (refs 6 and 26). Indeed, we deduce visibilities $V_{1a} = (70 \pm 2)\%$ and $V_{1b} = (71 \pm 2)\%$ for the associated conditional states.

Table 1 | Diagonal elements of the density matrix $\tilde{\rho}_{2L,2R}$, deduced from the records of photo-electric counts

Probability	D_{1a}	D_{1b}
p_{00}	0.98510 ± 0.00007	0.98501 ± 0.00007
p_{10}	$(7.38 \pm 0.05) \times 10^{-3}$	$(6.19 \pm 0.04) \times 10^{-3}$
p_{01}	$(7.51 \pm 0.05) \times 10^{-3}$	$(8.78 \pm 0.05) \times 10^{-3}$
p_{11}	$(1.7 \pm 0.2) \times 10^{-5}$	$(1.9 \pm 0.2) \times 10^{-5}$

The values of p_{ij} are referenced to the location of detectors $D_{2a,2b,2c}$, and were obtained by considering unit detection efficiency, which gives a more conservative (smaller) lower bound for the concurrence than the actual (larger) field concurrence for finite efficiency < 1 . See the Supplementary Information for further details, and equation (3).

A notable feature of these results is that the interference fringes have relative phase π for the cases of detection at D_{1a}, D_{1b} , in agreement with equations (1) and (2). We observe similar fringes if the phase η_1 between the write beams is varied instead of φ . Moreover, if the fields $1_L, 1_R$ are combined at the beamsplitter BS_1 with orthogonal polarizations (by way of the half-wave plate in Fig. 1a), we find that the visibility from interference of fields $2_L, 2_R$ drops to near zero, because in this case there is no longer measurement-induced entanglement associated with quantum interference for detection of fields $1_L, 1_R$ (see Supplementary Information).

With equation (4), the measured values for the visibility V and for the various p_{ij} are sufficient to deduce a lower bound for the concurrence C for the field state $\tilde{\rho}_{2L,2R}$ at the location of detectors $D_{2a,2b,2c}$. With no correction for detection efficiencies or propagation losses, and without subtraction of any background, we find:

$$C_{1a}(\tilde{\rho}_{2L,2R}) = (2.4 \pm 0.6) \times 10^{-3} > 0, \quad (5)$$

$$C_{1b}(\tilde{\rho}_{2L,2R}) = (1.9 \pm 0.6) \times 10^{-3} > 0$$

conditioned upon detection at either D_{1a} or D_{1b} . This conclusively demonstrates a non-zero degree of entanglement between the ensembles, albeit with the concurrence $C_{L,R}$ small. The small difference between the concurrence for the states conditioned on D_{1a} or D_{1b} can be explained by an asymmetry in BS_1 (see Supplementary Information).

Beyond the firm lower bound given by equation (5), we can make a better estimate of the degree of entanglement $C_{L,R}$ between the L and R ensembles by using detailed measurements of the propagation

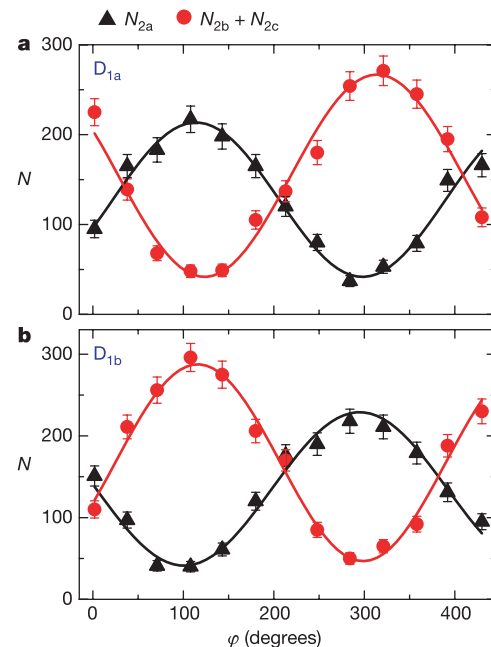


Figure 2 | Coherence between the atomic ensembles L, R induced by a measurement event of the fields 1_L and 1_R at detector D_{1a} or D_{1b} . Shown is the number of coincidences N_{2a} (triangles) and $N_{2b} + N_{2c}$ (circles) recorded by the respective detectors $D_{2a,2b,2c}$ for the fields 2_L and 2_R with the interferometer arrangement of Fig. 1b as a function of the relative phase φ . In **a**, $N_{2a,2b,2c}$ are conditioned upon a detection event at D_{1a} with no count at D_{1b} , while in **b**, $N_{2a,2b,2c}$ are conditioned upon an event at D_{1b} with no count at D_{1a} . At each setting of φ , data are acquired for 150 s with a detection window of width 190 ns. Although the interference fringes have comparable visibility, the different sizes arise from unequal quantum efficiencies for detectors D_{2a} and $D_{2b,2c}$ (see Supplementary Information). The visibility values are obtained from an average of the visibilities of the red and black curves, respectively. Error bars reflect $\pm$ one standard deviation due to the finite number of counts.

efficiencies from the atomic ensembles to the plane z_0 of the detectors shown in Fig. 1b (see Supplementary Information). Figure 3 gives an inference of the density matrix $\hat{\rho}_{2_L, 2_R}^{z_i}$ and thereby of the concurrence $C^{z_i}(\hat{\rho}_{2_L, 2_R}^{z_i})$ at z_0 and at two other locations $z_{i=1,2}$ along the path from the ensembles to the detectors (see Fig. 1b), assuming a constant visibility. In general, C increases in direct correspondence to the reduced level of losses for the 2_L and 2_R fields at locations closer to the ensembles. At location z_2 corresponding to the output edges of the atomic ensembles, we find the result:

$$C_{L,R}^{1a} \geq C_{1a}^{z_2}(\hat{\rho}_{2_L, 2_R}^{z_2}) \approx 0.021 \pm 0.006 > 0,$$

$$C_{L,R}^{1b} \geq C_{1b}^{z_2}(\hat{\rho}_{2_L, 2_R}^{z_2}) \approx 0.016 \pm 0.006 > 0$$
(6)

To move beyond this result, we need more detailed information about the efficiencies $\xi_{L,R}$ with which stored excitation in the atomic ensembles is converted to the propagating light fields 2_L and 2_R . Our earlier measurements included comparisons to a simple model²¹ and

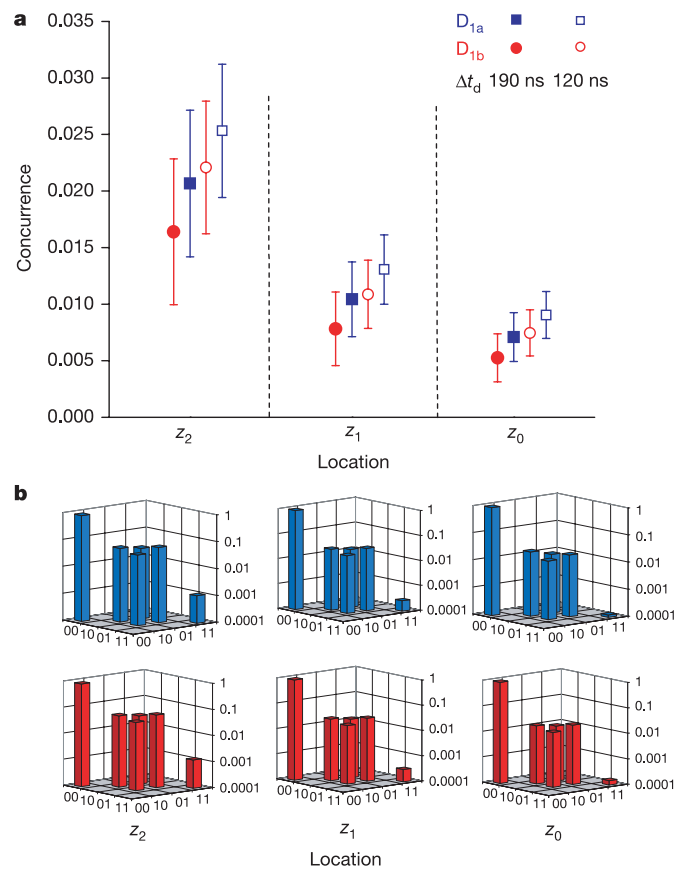


Figure 3 | Inference of the concurrence C^{z_i} (a) and density matrix $\hat{\rho}_{2_L, 2_R}^{z_i}$ (b) at the three locations z_i indicated in Fig. 1b. At each location, two pairs of results are given corresponding to the measurement-induced state created from detection at D_{1a} and D_{1b} , taking into account the efficiency of the detectors and propagation losses. **a**, Concurrence C , for two different detection windows Δt_d at $D_{2a, 2b, 2c}$. Filled symbols are for $\Delta t_d = 190$ ns, enough to acquire the whole temporal wavepacket of field 2. Open symbols are for $\Delta t_d = 120$ ns. We see then that the degree of entanglement can be further enhanced, similar to the increase of nonclassical correlations between fields 1 and 2 reported in ref. 23 for specific detection windows for these fields. All values shown in this figure, including the ones for z_0 , are already corrected for the efficiencies of the detectors. Error bars reflect ± 1 standard deviation, taking into account the finite number of counts and the uncertainties in the efficiency and propagation loss. **b**, Density matrix $\hat{\rho}_{2_L, 2_R}^{z_i}$ given in the basis $|n\rangle_{2_L}, |m\rangle_{2_R}$ corresponding to equation (3) with $\{n, m\} = \{0, 1\}$ for $\Delta t_d = 190$ ns.

allowed an inference $\xi_{L,R} \approx 0.10 \pm 0.05$. The measurement of the losses together with the values of p_{ij} at the detectors yields $p_{10} + p_{01} \approx 11\%$ at the output of the ensembles (z_2 plane) for our current experiment. This value together with the estimated $\xi_{L,R}$ then indicates that $p_{00} \rightarrow 0$ for the conditional state $\rho_{L,R}$ of the ensembles, so that $C_{L,R} \approx V \approx 0.7$, suggesting that $\rho_{L,R}$ is close to the ideal entangled state of equation (1). The low measured values for the entanglement between fields 2_L and 2_R are apparently principally a consequence of the low readout efficiency $\xi_{L,R}$ of the atomic excitation. We stress that this inference of C for the state inside the ensembles must be confirmed by subsequent experiments and is offered here to provide some insight into future prospects for quantum protocols with entangled ensembles. This also emphasizes that a central point in subsequent work should be the improvement of $\xi_{L,R}$.

In conclusion, we have achieved entanglement between a pair of atomic ensembles separated by 2.8 m, with the entangled state involving one spin excitation within a collective system of roughly 10^5 atoms at each site L and R. The entangled state is generated by and conditioned upon an initial detection event, and is thus produced in a probabilistic fashion. However, this initial event heralds unambiguously the creation of an entangled state between L and R ensembles, which is physically available for subsequent use, as, for example, by mapping to propagating optical fields, which can in principle be accomplished with high efficiency. We emphasize that our measurements relate to an actual physical state of the L and R ensembles and of the 2_L and 2_R fields, and are not an inference of a state based upon post-selection. Our work provides the first example of a stored atomic entangled state that can be transferred to entangled light fields, and significantly extends laboratory capabilities for entanglement generation, with now-entangled states of matter stored with separation a thousand-fold larger than was heretofore possible for qubits. With our current set-up, we have demonstrated $\Delta t_s \approx 1 \mu\text{s}$ for storing entanglement. However, this should readily be extended to $\Delta t_s \approx 10 \mu\text{s}$, and new trapping schemes have the potential to lead to $\Delta t_s \approx 1$ s (ref. 17). The distance scale for separating the L and R ensembles is limited by the length $l_0 \approx 2$ km for fibre optic attenuation at our write wavelength of 852 nm. Extensions to scalable quantum networks over larger distances will require the realization of a quantum repeater⁶, for which we have now laid the essential foundation.

METHODS

Atomic ensembles and optical pulses. Each of the L and R atomic ensembles is obtained from caesium atoms in a magneto-optical trap (MOT)^{17,21}. Measurements are carried out in a cyclic fashion consisting first of a period of cooling and trapping to form the MOT, followed by an interval during which the magnetic fields for the MOT are switched off. After waiting 3 ms for the magnetic field to decay¹⁷, we initiate a sequence of measurement trials, where for each trial the atoms are initially prepared in level $|g\rangle$. The write pulse is at 852 nm, with a duration of 150 ns and is detuned 10 MHz below the $|g\rangle \rightarrow |e\rangle$ transition. The read pulse is at 894 nm, with a duration of 130 ns and is resonant with the $|s\rangle \rightarrow |e'\rangle$ transition. At the end of each trial, the sample is pumped back to level $|g\rangle$ by illuminating the atomic cloud with trapping and repumping light for 0.7 μs and 1 μs respectively, and then a new trial is initiated with period of 3 μs . The total duration for a sequence of measurement trials is 5 ms, after which the measurement interval is terminated and a new MOT is formed in preparation for the next sequence of trials at a rate of 40 Hz.

Received 31 August; accepted 19 October 2005.

- Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, UK, 2000).
- Quantum Information and Computation Roadmap (http://qist.lanl.gov/qcomp_map.shtml) (2004).
- Copsey, D. et al. Toward a scalable, silicon-based quantum computing architecture. *IEEE J. Selected Topics Quant. Electron.* **9**, 1552–1569 (2003).
- Briegleb, H.-J., Dür, W., Cirac, J. I. & Zoller, P. Quantum repeaters: the role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).

5. Giovannetti, G., Lloyd, S. & Maccone, L. Quantum-enhanced measurements: beating the standard quantum limit. *Science* **306**, 1330–1336 (2004).
6. Duan, L.-M., Lukin, M., Cirac, J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414**, 413–418 (2001).
7. Clauser, J. F. & Shimony, A. Bell's theorem: experimental tests and implications. *Rep. Prog. Phys.* **41**, 1881–1927 (1978).
8. Aspect, A., Dalibard, J. & Roger, G. Experimental test of Bell's inequalities using time-varying analyzers. *Phys. Rev. Lett.* **49**, 1804–1807 (1982).
9. Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics* Ch. 12 (Cambridge Univ. Press, New York, 1995).
10. Haffner, H. *et al.* Robust entanglement. *Appl. Phys. B* **81**, 151–153 (2005).
11. Langer, C. *et al.* Long-lived qubit memory using atomic ions. *Phys. Rev. Lett.* **95**, 060502 (2005).
12. Julsgaard, B., Kozhekin, A. & Polzik, E. S. Experimental long-lived entanglement of two macroscopic objects. *Nature* **413**, 400–403 (2001).
13. Marcikic, I. *et al.* Distribution of time-bin entangled qubits over 50 km of optical fiber. *Phys. Rev. Lett.* **93**, 180502 (2004).
14. Ou, Z. Y., Pereira, S. F., Kimble, H. J. & Peng, K.-C. Realization of the Einstein–Podolsky–Rosen paradox for continuous variables. *Phys. Rev. Lett.* **68**, 3663–3666 (1992).
15. Turchette, Q. A. *et al.* Deterministic entanglement of two trapped ions. *Phys. Rev. Lett.* **81**, 3631–3634 (1998).
16. Blinov, B. B., Moehring, D. L., Duan, L.-M. & Monroe, C. Observation of entanglement between a single-trapped atom and a single photon. *Nature* **428**, 153–157 (2004).
17. Felinto, D., Chou, C. W., de Riedmatten, H., Polyakov, S. V. & Kimble, H. J. Control of decoherence in the generation of photon pairs from atomic ensembles. *Phys. Rev. A* **72**, 053809 (2005).
18. Kuzmich, A. *et al.* Generation of nonclassical photon pairs for scalable quantum communication with atomic ensembles. *Nature* **423**, 731–734 (2003).
19. van der Wal, C. H. *et al.* Atomic memory for correlated photon states. *Science* **301**, 196–200 (2003).
20. Jiang, W., Han, C., Xue, P., Duan, L.-M. & Guo, G.-C. Nonclassical photon pairs generated from a room-temperature atomic ensemble. *Phys. Rev. A* **69**, 043819 (2004).
21. Chou, C. W., Polyakov, S. V., Kuzmich, A. & Kimble, H. J. Single-photon generation from stored excitation in an atomic ensemble. *Phys. Rev. Lett.* **92**, 213601 (2004).
22. Eisaman, M. D. *et al.* Shaping quantum pulses of light via coherent atomic memory. *Phys. Rev. Lett.* **93**, 233602 (2004).
23. Polyakov, S. V., Chou, C. W., Felinto, D. & Kimble, H. J. Temporal dynamics of photon pairs generated by an atomic ensemble. *Phys. Rev. Lett.* **93**, 263601 (2004).
24. Balic, V., Braje, D. A., Kolchin, P., Yin, G. Y. & Harris, S. E. Generation of paired photons with controllable waveforms. *Phys. Rev. Lett.* **94**, 183601 (2005).
25. Matsukevich, D. N. *et al.* Entanglement of a photon and a collective atomic excitation. *Phys. Rev. Lett.* **95**, 040405 (2005).
26. Matsukevich, D. N. & Kuzmich, A. Quantum state transfer between matter and light. *Science* **306**, 663–666 (2004).
27. van Enk, S. & Kimble, H. J. Comment on 'Quantum state transfer between matter and light'. *Science* **309**, 1187b (2005).
28. Matsukevich, D. N. & Kuzmich, A. Response to comment on 'Quantum state transfer between matter and light'. *Science* **309**, 1187c (2005).
29. Duan, L.-M., Cirac, J. I. & Zoller, P. Three-dimensional theory for interaction between atomic ensembles and free-space light. *Phys. Rev. A* **66**, 023818 (2002).
30. Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80**, 2245–2248 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We gratefully acknowledge J. Hall and J. Ye for discussions about phase stabilization. This research is supported by the Advanced Research and Development Activity (ARDA), by the National Science Foundation, and by the Caltech MURI Center for Quantum Networks. D.F. acknowledges financial support by CNPq (Brazilian agency). H.d.R. acknowledges financial support by the Swiss National Science Foundation. S.J.v.E. thanks L. Huelsbergen for assistance in computer matters.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.J.K. (hjkimble@caltech.edu).

Storage and retrieval of single photons transmitted between remote quantum memories

T. Chanelière¹, D. N. Matsukevich¹, S. D. Jenkins¹, S.-Y. Lan¹, T. A. B. Kennedy¹ & A. Kuzmich¹

An elementary quantum network operation involves storing a qubit state in an atomic quantum memory node, and then retrieving and transporting the information through a single photon excitation to a remote quantum memory node for further storage or analysis. Implementations of quantum network operations are thus conditioned on the ability to realize matter-to-light and/or light-to-matter quantum state mappings. Here we report the generation, transmission, storage and retrieval of single quanta using two remote atomic ensembles. A single photon is generated from a cold atomic ensemble at one site¹, and is directed to another site through 100 metres of optical fibre. The photon is then converted into a single collective atomic excitation using a dark-state polariton approach². After a programmable storage time, the atomic excitation is converted back into a single photon. This is demonstrated experimentally, for a storage time of 0.5 microseconds, by measurement of an anti-correlation parameter. Storage times exceeding ten microseconds are observed by intensity cross-correlation measurements. This storage period is two orders of magnitude longer than the time required to achieve conversion between photonic and atomic quanta. The controlled transfer of single quanta between remote quantum memories constitutes an important step towards distributed quantum networks.

A quantum network, consisting of quantum nodes and interconnecting channels, is an outstanding goal of quantum information science. Such a network could be used for distributed computing or for the secure sharing of information between spatially remote parties^{1,3–7}. While it is natural that the network's fixed nodes (quantum memory elements) could be implemented by using matter in the form of individual atoms or atomic ensembles, it is equally natural that light fields be used as carriers of quantum information (flying qubits) using optical fibre interconnects. The matter–light interface seems inevitable since the local storage capability of ground state atomic matter cannot be easily recreated with light fields. Interfacing material quanta and single photons is therefore a basic primitive of a quantum network.

The potential of atomic ensembles to serve as quantum memories has recently attracted considerable attention^{1,2,8–11}, spawning two distinct lines of research. In one, using the physics of 'slow light' propagation in an optically thick atomic ensemble, weak coherent laser pulses have been stopped and retrieved in a controlled fashion^{2,12–14}. In the other, motivated by the seminal proposal of Duan, Lukin, Cirac and Zoller (DLCZ)¹, correlated pairs of photons and single photons have been produced from an atomic ensemble^{15–20}. Collective atomic qubits, atom–photon entanglement, and quantum state transfer from atomic to photonic qubits have also been demonstrated²¹. These initial experimental demonstrations within the DLCZ paradigm were beset by short atomic coherence times, of the order of the laser pulse length. In contrast, recent advances in atomic ensemble research²² allow for long quantum

memory times, in excess of ten microseconds in the present work, more than two orders of magnitude longer than the duration of the laser pulses involved in the protocols.

Here we report the synthesis of these two lines of research by demonstrating the generation, transmission, storage and retrieval of single photons using remote atomic ensembles as quantum memories. The essential ingredient that enables the completion of this synthesis, and which we report here, is the ability to convert single photons into single collective atomic excitations. In our experiment the remote quantum memories are based on cold atomic clouds of ⁸⁵Rb confined in magneto-optical traps (MOTs) at Sites A and B, as shown in Fig. 1. Sites A and B are physically located in adjacent laboratories, with a 100-metre-long single-mode optical fibre serving as the quantum information channel.

Our protocol begins with the generation of single photons at Site

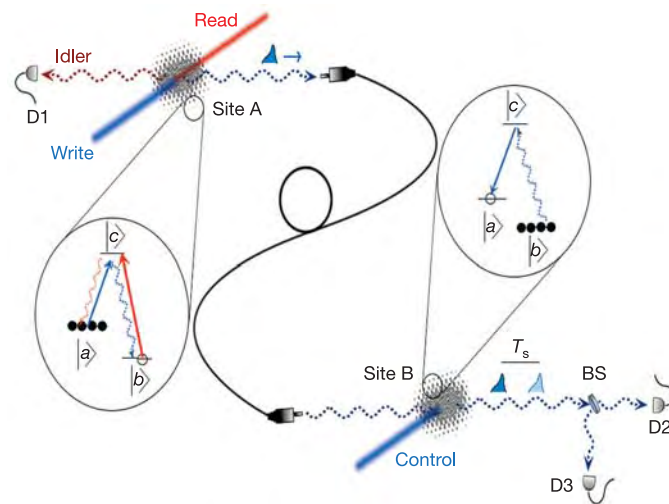


Figure 1 | A schematic diagram of our experimental set-up, demonstrating generation, transmission, storage and retrieval of single photon excitations of the electromagnetic field. Two atomic ensembles at Sites A and B are connected by a single-mode fibre. The insets show the structure and the initial populations of atomic levels for the two ensembles. All the light fields responsible for trapping and cooling, as well as the quadrupole magnetic fields in both MOTs, are shut off during the period of the protocol. The ambient magnetic field at each site is compensated by three pairs of Helmholtz coils (not shown). Correlated signal and idler fields are generated at Site A. The signal field is transmitted via optical fibre from Site A to Site B, where it is converted to atomic excitation, stored for a duration T_s , and subsequently retrieved. A Hanbury Brown-Twiss set-up consisting of a beamsplitter BS and two detectors D2 and D3, together with detector D1 for the idler field, are used to verify the single photon character of the retrieved field.

¹School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA.

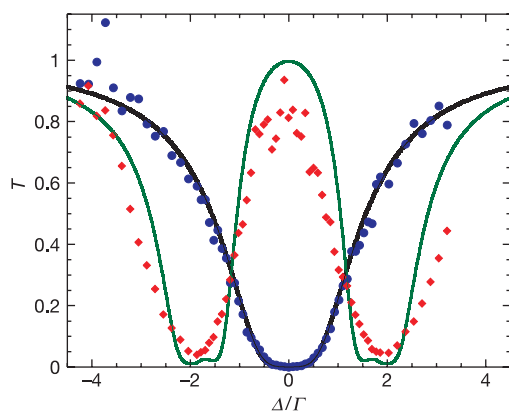


Figure 2 | Measured transmission spectra of a coherent probe field as a function of probe detuning in the presence of, and absence of, EIT. Data are taken using 700-ns-long coherent laser pulses. T is the intensity transmittance, Δ is the probe detuning and Γ is the decay rate of level $|c\rangle$. In the absence of control field (circles) the probe is strongly absorbed near resonance, whereas with the control field on (diamonds) the medium becomes transparent. Each probe pulse contains on average 0.3 photons. Each data point is an average of 2×10^5 experimental trials. The optical thickness $d = 8$ and the control field Rabi frequency $\Omega = 3\Gamma$ are used to obtain the solid curves, based on the theoretical model discussed in the Supplementary Information.

A, using an improved version of the DLCZ approach in the off-axis, counter-propagating geometry^{20,22}. The fibre channel directs the signal field to Site B where an optically thick atomic ensemble is prepared in level $|b\rangle$ (right inset in Fig. 1). The signal field propagation in the atomic medium is controlled by an additional laser field ('control') through the process of electromagnetically induced transparency (EIT)^{23,24}. As we deal with an unpolarized atomic ensemble, we must take into account the Zeeman degeneracy of the atomic levels. Choosing the same circular polarizations for both the probe and the control fields allows us to retain transparency, as discussed in more detail in the Supplementary Information. In Fig. 2 we show the EIT transmission spectrum recorded for a coherent laser probe field instead of the signal field. Evidently, in the absence of the control light the probe field is absorbed by the optically thick sample. With the addition of the c.w. control field, the medium is rendered transparent around the $|b\rangle \leftrightarrow |c\rangle$ transition resonance $\Delta = 0$.

The control field strongly modifies the group velocity of the signal field. For a time-dependent control field, a strong reduction of the group velocity of the propagating signal field can be understood in terms of a coupled matter–light field excitation known as a 'dark-state polariton'. By adiabatically switching off the control field, the coupled excitation can be converted into a pure atomic excitation, that is, the signal field is 'stopped'^{2,13,14}. An important condition to achieve storage is a sufficiently large optical thickness of the atomic sample, which enables strong spatial compression of the incident signal field⁹. In our experiment the measured optical thickness $d \approx 8$. Figure 3 compares our observations with the predictions of a theoretical model. Figure 3a compares the propagation of the signal pulse in vacuum and in the atomic medium under conditions of EIT with a c.w. control field. The observed pulse delay under conditions of EIT is about 20 ns, corresponding to more than three orders of magnitude reduction in group velocity. Figure 3b shows the effect of turning off the control-storage field when the signal pulse is approximately centred in the medium, and the subsequent retrieval of the signal field when the control-retrieval field is switched back on after a 500 ns storage time. Figure 3c shows retrieval after a storage time of 15 μ s. Qualitative agreement of the pulse shapes has been obtained in our theoretical analysis of the protocol using the full Zeeman structure of the atoms and a classical description of the signal field (Fig. 3d–f).

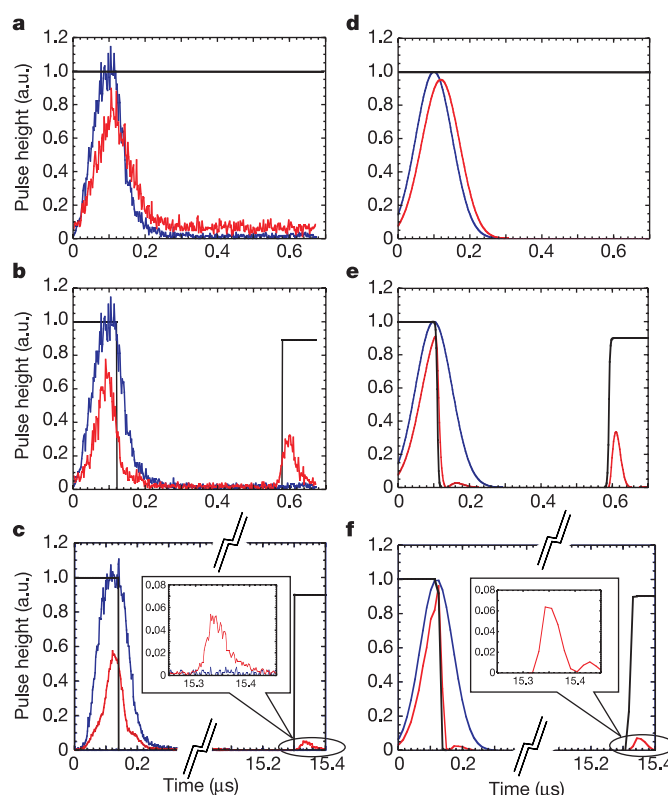


Figure 3 | Experimental and theoretical pulse shapes as a function of time, showing EIT, storage and retrieval. The colour code is: control field, black; pulse in vacuum, blue; delayed, stored and retrieved field, red. Panel a with a c.w. control field shows EIT pulse delay. In panel b the control field is switched off and then on again after 500 ns, and the panel shows light storage and retrieval. Panel c is similar to b but with a 15 μ s storage. Panels d, e and f are corresponding theoretical plots.

In order to verify the single-photon character of the signal field (1) without storage, and (2) with storage and retrieval, we use a Hanbury Brown-Twiss detection scheme, employing a beamsplitter followed by two single photon counters, as shown in Fig. 1²⁵. To provide such characterization, we note that classical fields must satisfy a criterion $\alpha \geq 1$ based on the Cauchy-Schwarz inequality^{25,26}. For an ideally prepared single photon state $\alpha \rightarrow 0$. Here the anticorrelation parameter α is a function of the storage time T_s , and is given by the ratio of various photoelectric detection probabilities which are measured by the set of detectors D1, D2 and D3 (see Methods):

$$\alpha(T_s) = \frac{p_1 p_{123}}{p_{12} p_{13}} \quad (1)$$

As an auxiliary measure of signal-idler field correlations, and as a way to quantify the quantum memory storage time, we also evaluate the normalized intensity cross-correlation function $g_{si} \equiv (p_{12} + p_{13}) / [p_1(p_2 + p_3)]$ (ref. 27). In particular, it serves to estimate the total efficiency and background levels in the experiment, since g_{si} is, by definition, independent of efficiencies whereas p_1 is proportional to the overall idler channel efficiency.

First we measure g_{si} and α without storage at Site B (that is, with no atomic sample in place), and the results are displayed in Fig. 4a and b, respectively. Next we add an optically thick atomic sample at Site B, and perform storage of duration $T_s = 500$ ns and subsequent retrieval of the signal field, with results shown in Fig. 4c and d, respectively. No correction for background or dark counts were made to any of the experimental counting rates. The curve fits of g_{si} are based on a simple theoretical model, and allow us to obtain the efficiency in the idler channel and the background contributions to p_2 and p_3 for the stored signal field. These same values are used to

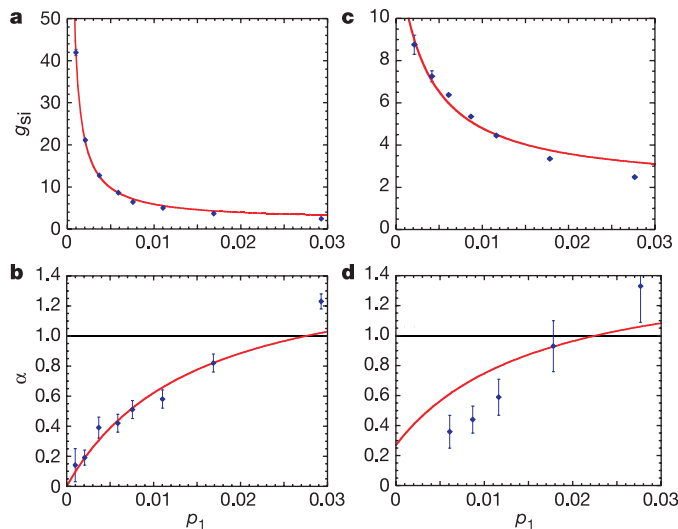


Figure 4 | Measured intensity cross-correlation function g_{si} and anticorrelation function α as a function of the idler photoelectric detection probability p_1 . Panels **a** and **b** are for the source (propagation in vacuum). Panels **c** and **d** are for stopped, stored for 500 ns, and retrieved signal field. The solid lines are based on a theoretical model that includes losses and background. Error bars represent $\pm$ one standard deviation, and are based on the statistics of the photoelectric counting events.

produce the corresponding theoretical curves in Fig. 4b and d. The measured values of $\alpha < 1$, displayed in Fig. 4b and d, confirm the single-photon character of both the source and retrieved signal fields (with the minimum values of $\alpha = 0.14 \pm 0.11$ and $\alpha = 0.36 \pm 0.11$, respectively). Overall, we estimate that the probability p_s for successful generation, transmission, storage, retrieval and detection of a signal photon is approximately $p_s \approx 10^{-5}$ for each trial. The efficiency of photon storage and retrieval E can be estimated as the ratio of the values of $p_2 + p_3$ with and without storage. We find $E \approx 0.06$, in agreement with the theoretical result shown in Fig. 3e.

To investigate the storage capability of our quantum memory at Site B, we measure g_{si} as a function of the storage time of the signal field T_s (Fig. 5). A gaussian fit provides a time constant $\tau = 11 \mu\text{s}$, which is an estimate of our quantum memory time. The collapse is consistent with the Larmor precession of a dark-state polariton in an unpolarized atomic ensemble in a residual magnetic field. Experimentally we attempt to null the uniform, d.c. component of the magnetic field. A definitive way to distinguish whether the collapse is due to uniform or non-uniform and a.c. fields is to measure the damping time of the periodic revivals of the retrieved signal field at longer storage times. In a uniform magnetic field, undamped revivals of the dark-state polariton should occur at times equal to nT_L , where T_L is the Larmor period for level $|a\rangle$ or $|b\rangle$ and n can be either integer or half-integer, depending on the direction of the magnetic field relative to the light beam geometry (a synopsis of these ideas is given in the Supplementary Information, with the full theory presented in ref. 28). We have conducted separate experiments with an externally applied magnetic field²⁹, which suggest that the collapse in the present experiment is probably due to magnetic field gradients and/or a.c. fields at the level of a few tens of mG. However, more extensive investigations to quantitatively determine the temporal and spatial structure of the residual magnetic field, and the various contributions to it, are ongoing.

We have demonstrated generation, storage and retrieval of single quanta transmitted between two remote atomic ensembles serving as quantum memory elements. The control of the matter-field interface at the level of single quanta, and at remote sites, is encouraging for further developments and applications in quantum information science. In particular, the storage of a photonic qubit, with two

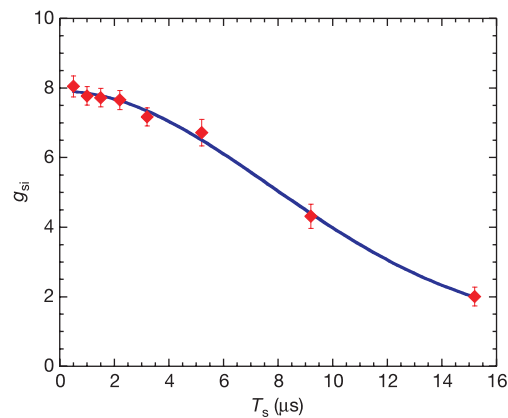


Figure 5 | Normalized signal-idler intensity correlation function g_{si} as a function of the storage time T_s at Site B. Data (diamonds) are taken for $p_1 = 0.0047$, but with a smaller background contribution than that of Fig. 4c and d. The full curve is a fit of the form $1 + B\exp(-t^2/\tau^2)$ with $B = 7$ and the collapse time $\tau = 11 \mu\text{s}$ as adjustable parameters. Error bars represent $\pm$ one standard deviation, and are based on the statistics of the photoelectric counting events.

logical states, would represent a crucial advance. In order to achieve this, the quantum memory at Site B would likewise need a second logical state, so as to realize a collective atomic qubit. Two different approaches for such qubits have already been demonstrated^{21,22}. If a second logical state were added to both quantum memories at Sites A and B, generation of remote entanglement of two atomic qubits would be possible.

METHODS

To generate single photons at Site A, we use the DLCZ approach in the off-axis, counter-propagating geometry introduced by Harris and co-workers²⁰. The insets in Fig. 1 indicate schematically the structure of the three atomic levels involved, $|a\rangle$, $|b\rangle$ and $|c\rangle$, where $\{|a\rangle; |b\rangle\}$ correspond to the $5S_{1/2}$, $F = \{3, 2\}$ levels of ^{85}Rb , and $|c\rangle$ represents the $\{5P_{1/2}, F = 3\}$ level associated with the D_1 line at 795 nm. The experimental sequence begins with an unpolarized sample of atoms prepared in level $|a\rangle$ (left inset of Fig. 1). A 160-ns-long write laser pulse tuned to the $|a\rangle \rightarrow |c\rangle$ transition is focused into the MOT with a gaussian waist of about 400 μm . The write pulse generates a cone of forward Raman-scattered signal field via the $|c\rangle \rightarrow |b\rangle$ transition. We collect a gaussian mode centred around the momentum $\mathbf{k}_s$ that forms an angle of about 2° with the write beam. The write pulse is so weak that on average less than one photon is scattered into the collected mode for each pulse. The signal field is coupled into the 100-m-long fibre connecting Sites A and B.

For each signal photon emission event, a correlated collective atomic excitation is created in the atomic ensemble. After a delay $\Delta t = 200$ ns, a 140-ns-long counter-propagating read laser pulse resonant with the $|b\rangle \rightarrow |c\rangle$ transition illuminates the atomic ensemble and converts the atomic excitation into the idler field. Under the conditions of collective enhancement, the idler field is emitted with high probability into the mode determined by the phase-matching condition $\mathbf{k}_i = \mathbf{k}_w + \mathbf{k}_r - \mathbf{k}_s$, where $\mathbf{k}_i$, $\mathbf{k}_w$ and $\mathbf{k}_r$ are the wave vectors of the idler, write and read fields, respectively. The waist of the signal-idler mode in the MOT is about 150 μm . The idler field is directed onto a single photon counter D1. Ideally, photoelectric detection of the idler field projects the quantum state of the signal field into a single photon state. The repetition rate of the experiment is $2 \times 10^5 \text{ s}^{-1}$. Each data point in Fig. 4 involves an average over a time period that varied from several minutes up to 1.5 h for the data point with the lowest value of p_1 in d.

To measure the photoelectric detection probabilities $p_1, p_2, p_3, p_{13}, p_{12}, p_{23}$ and p_{123} , the outputs of the detectors are fed to three 'Stop' inputs of the time-interval analyser, which records the arrival times with a 2 ns time resolution. The electronic pulses from the detectors D1, D2, D3 are gated for periods $[t_0^i, t_0^i + T_g^i]$, with $T_g^1 = 140$ ns, $T_g^2 = T_g^3 = 240$ ns, respectively, centred on the times determined by the write and read (for no storage) or control-retrieval (for storage) laser pulses. Counts recorded outside the gating periods are therefore removed from the analysis. The list of recorded events allows us to determine the single-channel photoelectric event probabilities $p_i = N_i/M$, where

N_i is the total number of counts in the i th channel and M is the number of experimental trials (for D_i , $i = 1, 2, 3$). If photoelectric detections in different channels i, k, m happen within the same gating period, they contribute to the corresponding joint probabilities $p_{ij} = N_{ij}/M$, where N_{ij} is the total number of coincidences between D_i and D_j , where $i, j = 1, 2, 3$. The joint probability of all three detectors registering a count is given by $p_{123} = N_{123}/M$.

Received 23 August; accepted 13 October 2005.

1. Duan, L.-M., Lukin, M., Cirac, J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414**, 413–418 (2001).
2. Fleischhauer, M. & Lukin, M. D. Dark-state polaritons in electromagnetically induced transparency. *Phys. Rev. Lett.* **84**, 5094–5097 (2000).
3. Ekert, A. K. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* **67**, 661–663 (1991).
4. Bouwmeester, D. *et al.* Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
5. Boschi, D. *et al.* Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **80**, 1121–1125 (1998).
6. Briegel, H.-J., Duer, W., Cirac, J. I. & Zoller, P. Quantum repeaters: The role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
7. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–52 (2001).
8. Raymer, M. G. & Walmsley, I. A. in *Progress in Optics* Vol. XXVIII (ed. Wolf, E.) 181–270 (North-Holland, Amsterdam, 1996).
9. Lukin, M. D. Colloquium: Trapping and manipulating photon states in atomic ensembles. *Rev. Mod. Phys.* **75**, 457–472 (2003).
10. Kuzmich, A. & Kennedy, T. A. B. Non-symmetric entanglement of atomic ensembles. *Phys. Rev. Lett.* **92**, 030407 (2004).
11. Julsgaard, B. *et al.* Experimental demonstration of quantum memory for light. *Nature* **432**, 482–486 (2004).
12. Hau, L. V., Harris, S. E., Dutton, Z. & Behroozi, C. H. Light speed reduction to 17 metres per second in an ultracold atomic gas. *Nature* **397**, 594–598 (1999).
13. Phillips, D. F. *et al.* Storage of light in atomic vapor. *Phys. Rev. Lett.* **86**, 783–786 (2001).
14. Liu, C., Dutton, Z., Behroozi, C. H. & Hau, L. V. Observation of coherent optical information storage in an atomic medium using halted light pulses. *Nature* **409**, 490–493 (2001).
15. Kuzmich, A. *et al.* Generation of nonclassical photon pairs for scalable quantum communication with atomic ensembles. *Nature* **423**, 731–734 (2003).
16. van der Wal, C. H. *et al.* Atomic memory for correlated photon states. *Science* **301**, 196–200 (2003).
17. Jiang, W. *et al.* Nonclassical photon pairs generated from a room-temperature atomic ensemble. *Phys. Rev. A* **69**, 043819 (2004).
18. Chou, C. W., Polyakov, S. V., Kuzmich, A. & Kimble, H. J. Single photon generation from stored excitation in an atomic ensemble. *Phys. Rev. Lett.* **92**, 213601 (2004).
19. Eisaman, M. D. *et al.* Shaping quantum pulses of light via coherent atomic memory. *Phys. Rev. Lett.* **93**, 233602 (2004).
20. Balic, V., Braje, D. A., Kolchin, P., Yin, G. Y. & Harris, S. E. Generation of paired photons with controllable waveforms. *Phys. Rev. Lett.* **94**, 183601 (2005).
21. Matsukevich, D. N. & Kuzmich, A. Quantum state transfer between matter and light. *Science* **306**, 663–666 (2004).
22. Matsukevich, D. N. *et al.* Entanglement of a photon and a collective atomic excitation. *Phys. Rev. Lett.* **95**, 040405 (2005).
23. Harris, S. E. Electromagnetically induced transparency. *Phys. Today* **50**, 36–42 (1997).
24. Scully, M. O. & Zubairy, M. S. *Quantum Optics* (Cambridge Univ. Press, Cambridge, 1997).
25. Grangier, P., Roger, G. & Aspect, A. Experimental evidence for a photon anticorrelation effect on a beam splitter: A new light on single-photon interferences. *Europhys. Lett.* **1**, 173–179 (1986).
26. U'Ren, A. B. *et al.* Characterization of the non-classical nature of conditionally prepared single photons. *Phys. Rev. A* **72**, R021802 (2005).
27. Clauser, J. F. Experimental distinction between the quantum and classical field-theoretic predictions for the photoelectric effect. *Phys. Rev. D* **9**, 853–860 (1974).
28. Jenkins, S. D. *et al.* Theory of dark state polariton collapses and revivals. *Phys. Rev. A* (submitted); preprint at (<http://arXiv.org/quant-ph/0511017>) (2005).
29. Matsukevich, D. N. *et al.* Observation of dark state polariton collapses and revivals. *Phys. Rev. Lett.* (submitted); preprint at (<http://arXiv.org/quant-ph/0511015>) (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by NASA, Office of Naval Research Young Investigator Program, National Science Foundation, Research Corporation, Alfred P. Sloan Foundation, and Cullen-Peck Chair. We thank M. S. Chapman for discussions and E. T. Neumann for experimental assistance.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.K. (alex.kuzmich@physics.gatech.edu).

Electromagnetically induced transparency with tunable single-photon pulses

M. D. Eisaman¹, A. André¹, F. Massou¹, M. Fleischhauer^{1,2,3}, A. S. Zibrov^{1,2,4} & M. D. Lukin¹

Techniques to facilitate controlled interactions between single photons and atoms are now being actively explored^{1–7}. These techniques are important for the practical realization of quantum networks, in which multiple memory nodes that utilize atoms for generation, storage and processing of quantum states are connected by single-photon transmission in optical fibres^{1,2}. One promising avenue for the realization of quantum networks involves the manipulation of quantum pulses of light in optically dense atomic ensembles using electromagnetically induced transparency (EIT, refs 8, 9). EIT is a coherent control technique that is widely used for controlling the propagation of classical, multi-photon light pulses^{10–14} in applications such as efficient nonlinear optics¹⁵. Here we demonstrate the use of EIT for the controllable generation, transmission and storage of single photons with tunable frequency, timing and bandwidth. We study the interaction of single photons produced in a ‘source’ ensemble of ⁸⁷Rb atoms at room temperature with another ‘target’ ensemble. This allows us to simultaneously probe the spectral and quantum statistical properties of narrow-bandwidth single-photon pulses, revealing that their quantum nature is preserved under EIT propagation and storage. We measure the time delay associated with the reduced group velocity of the single-photon pulses and report observations of their storage and retrieval.

The basic idea of our experiments is illustrated in Fig. 1a. Single photons are prepared in an ensemble of room-temperature ⁸⁷Rb atoms (called the ‘source ensemble’) by first creating a single spin excitation via Raman scattering combined with single-photon detection, and later converting this atomic excitation ‘on demand’ into a single photon propagating in an optical fibre^{3,4,16–21}. Successful preparation of the single-photon pulse is conditional on detecting a single Raman-scattered photon^{16,17}. The single photons are directed via an optical fibre to a second atomic ensemble (‘target ensemble’), where their controlled interaction with coherently driven atoms is studied by combining EIT-based high-resolution spectroscopy and photon-counting measurements.

We begin by describing our source of narrow-bandwidth, frequency-tunable single photons with properties matching those of narrow atomic resonances^{16,17}. As illustrated in Fig. 1a, the source ensemble is initially prepared in the ground state $|g\rangle$. Atomic spin excitations to the state $|s\rangle$ are produced via spontaneous Raman scattering, induced by a laser beam referred to as the write laser. In this process, correlated pairs of frequency-shifted photons (so-called Stokes photons) and flipped atomic spins are created (corresponding to atomic Raman transitions into the state $|s\rangle$). Energy and momentum conservation ensure that by detecting a Stokes photon emitted in a particular direction, the atomic ensemble is prepared in state with exactly one flipped spin quantum in a well-defined spin-wave mode. Conditioned upon detecting a single Stokes photon, the stored single spin-wave quantum is coherently converted into a single-photon

anti-Stokes pulse by applying a second near-resonant laser beam (retrieve laser) after a controllable delay time¹². The direction, bandwidth, and central frequency of the single-photon anti-Stokes pulse is determined by the direction, intensity and frequency of the retrieve laser¹⁷. Specifically, the retrieve laser controls the rate of retrieval and propagation of the anti-Stokes pulse, thereby controlling its duration, and consequently its bandwidth. The central frequency of the single-photon pulse differs from the frequency of the retrieve laser by a fixed amount given by the $|g\rangle$ – $|s\rangle$ atomic transition frequency. We study the photon-number fluctuations in the Stokes and the anti-Stokes pulses using a Hanbury-Brown-Twiss-type setup, which allows us to measure normalized correlation functions $g^{(2)}(x, y) \equiv \langle : \hat{n}_x \hat{n}_y : \rangle / \langle \hat{n}_x \rangle \langle \hat{n}_y \rangle$, where $\hat{n}_i$ denotes the photon-number operator for field i , and $::$ denotes operator normal ordering^{22,23}.

To quantify the properties of the single-photon source, the target ensemble was first removed from the beam path. Figure 2 shows a measurement of the photon-number fluctuations of the anti-Stokes field conditioned on detecting a single Stokes photon, as a function of the detection probability $p\eta_S$ in the Stokes channel. (Here p is the Raman excitation probability, and η_S is the overall Stokes channel transmission.) The function $g^{(2)}(AS||n_S = 1)$ (where n_S is the number of detected Stokes photons; see Fig. 2) represents a measure of the photon-number fluctuations in the anti-Stokes pulses. An ideal single-photon source has no photon-number fluctuations ($g^{(2)}(AS||n_S = 1) = 0$); for classical coherent states $g^{(2)}(AS||n_S = 1) = 1$. In Fig. 2, $p\eta_S$ is varied by changing p via the write laser intensity. As p becomes much smaller than unity, we observe substantial suppression of the conditional intensity fluctuations in the anti-Stokes pulses ($g^{(2)} = 0.3 \pm 0.2$ for $p\eta_S = 0.06$ and $\eta_S = 0.27$) compared to the classical limit of unity. Typical conversion efficiencies of atomic excitations into anti-Stokes photons are 8–15%. These observations are in good agreement with a simple theoretical model²⁴ that considers realistic losses and background photons. The presence of loss on the Stokes channel means that detection of a single Stokes photon can result in more than one atomic excitation. Upon retrieval, this results in the undesired emission of more than one anti-Stokes photon. Even in the presence of loss, one can obtain almost perfect preparation of an atomic state with a single excitation by ensuring that the Raman excitation probability p is much less than one. In this case, the probability of emitting two photons is suppressed by $p \ll 1$. This condition is satisfied when $p\eta_S \ll \eta_S$, in agreement with the experimental observations in Fig. 2.

We next consider the interaction of these non-classical anti-Stokes pulses with the optically dense target ensemble (Fig. 1). Usually such a medium simply absorbs the incoming light, reducing its intensity and destroying its quantum state. To restore transparency and control the light propagation, EIT is used. The essence of EIT,

¹Physics Department, Harvard University, ²Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA. ³Fachbereich Physik, Technische Universität Kaiserslautern, D-67663 Kaiserslautern, Germany. ⁴P. N. Lebedev Institute of Physics, Moscow, 117924, Russia.

illustrated on the right-hand side of Fig. 1a, is a strong coupling of an incident light pulse (anti-Stokes field in Fig. 1a) to a long-lived $|g\rangle\text{--}|s\rangle$ atomic coherence (spin wave), mediated by a coherent laser ('EIT control' laser). This control laser converts the incoming light pulse into a so-called 'dark' spin state, thereby eliminating dissipative absorption and substantially reducing its group velocity²⁵. Note that EIT is effective only within a narrow range of frequencies associated with the spectral transparency window, which occurs when the frequency difference between the incident pulse and the control laser matches the frequency of the spin coherence.

The main idea behind our experimental implementation is to match the bandwidth and the central frequency of our single-photon source to the EIT transparency resonance of the target ensemble by tuning, respectively, the retrieve and the control laser intensities and frequencies¹⁷. The relative detuning between the retrieve and EIT control lasers is carefully controlled via acousto-optic modulators. Figure 3a shows the conditional probability of detecting an anti-Stokes photon transmitted through the target ensemble, $\langle n \rangle(AS||n_S = 1)$, as a function of the two-photon detuning δ (the difference between the anti-Stokes/EIT control laser frequency difference, and the $|g\rangle\text{--}|s\rangle$ transition frequency). The clear resonance structure displays maximum transmission for $\delta = 0$. At this point, the central frequency of the single photons coincides with the EIT resonance window, resulting in a three-fold increase in transmission, which corresponds to 60% transmission of the incident pulse. The observed conditional probabilities can be used to

quantify the correlations between the Stokes and anti-Stokes photon numbers using the normalized correlation function $R \equiv g^{(2)}(S, AS)^2 / g^{(2)}(S, S) g^{(2)}(AS, AS)$. Classical fields must obey the Cauchy-Schwartz inequality $R \leq 1$; $R > 1$ indicates non-classical correlations²². For the data at $\delta = 0$, $R = 1.85 \pm 0.12$, including all background and dark counts; as δ is tuned away from zero in either direction, R approaches the classical limit of unity.

Figure 3b shows the normalized photon-number fluctuations for the transmitted anti-Stokes field conditioned upon detection of $n_S = 1$ Stokes photon, $g^{(2)}(AS||n_S = 1)$, versus δ . We observe that $g^{(2)}(AS||n_S = 1)$ retains its non-classical character upon transmission through the target ensemble only when near the centre of the EIT transparency window. The minimum measured value of $g^{(2)}(AS||n_S = 1)$, occurring at $\delta = 0$ (0.50 ± 0.14), is essentially equal to the value measured by removing the target ensemble from the beam path (0.51 ± 0.15 for the displayed set of data). It is important to emphasize that the maximum of $\langle n \rangle(AS||n_S = 1)$ and the minimum $g^{(2)}(AS||n_S = 1)$ both occur at $\delta = 0$. As δ is tuned away from zero in either direction, $\langle n \rangle(AS||n_S = 1)$ decreases while $g^{(2)}(AS||n_S = 1)$ approaches the classical limit of unity, indicating that the non-classical nature of the anti-Stokes pulse is preserved only within the EIT transparency window²⁶. The classical limit is also observed when the EIT control field is turned off. Likewise, $g^{(2)}(AS)$ obtained without conditioning exhibits no structure as a function of δ and again yields the classical limit of unity. Finally, we note that the photon-correlation data display a noise-enhancement feature on the

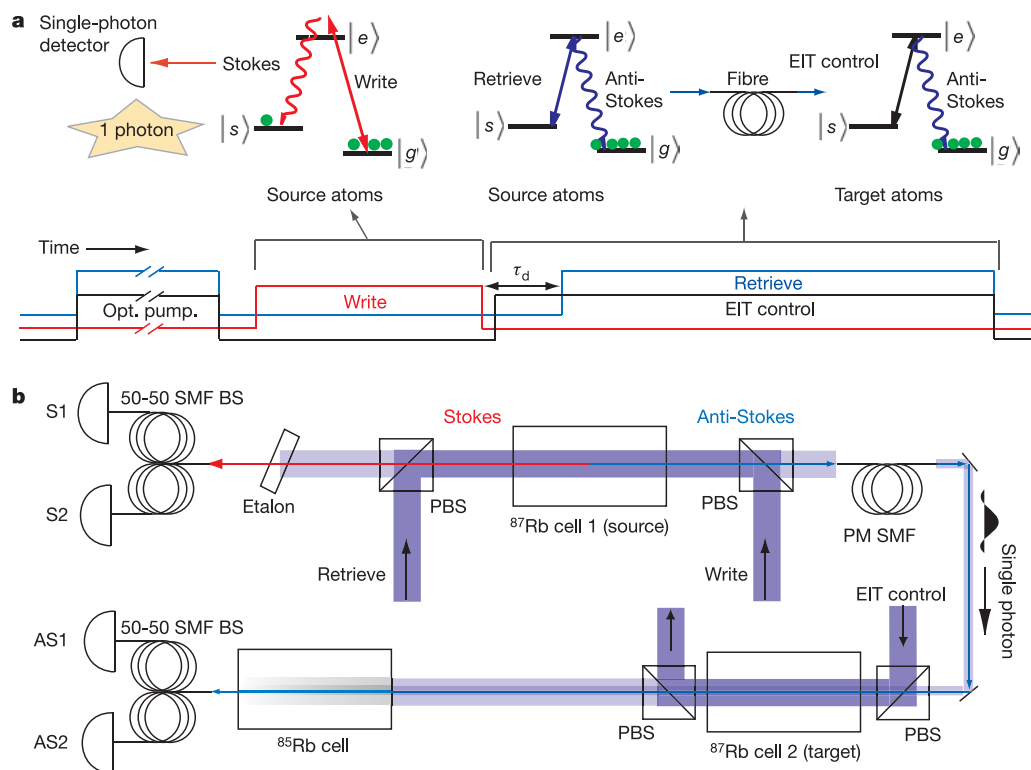


Figure 1 | Experimental procedure and set-up. **a**, Two ensembles of ^{87}Rb atoms are used, the 'source' and 'target' ensembles. In zero magnetic field, the atoms can be pictured as a three-level atom, with: $|g\rangle = |5^2S_{1/2}, F = 1\rangle$, $|s\rangle = |5^2S_{1/2}, F = 2\rangle$, and $|e\rangle$ corresponds to $|5^2P_{1/2}, F' = 1\rangle$ and $|5^2P_{1/2}, F' = 2\rangle$. The write laser and the retrieve laser couple respectively the $|g\rangle\text{--}|e\rangle$ and $|s\rangle\text{--}|e\rangle$ transition of the source atoms; the EIT control laser couples the $|s\rangle\text{--}|e\rangle$ transition of the target atoms. **b**, The write and retrieve lasers counter-propagate³⁰ inside the magnetically shielded source ensemble, and the EIT control laser and anti-Stokes field co-propagate inside the magnetically shielded target ensemble. The write and retrieve lasers have a diameter of 1 mm and 3 mm respectively at the centre of the source

ensemble. The single spatial mode defined by the detection fibres and optics has a diameter of $200\ \mu\text{m}$ at the centre of the source ensemble. The etalon is used to reflect the fraction of the write laser not filtered by the polarizing beamsplitters, and the ^{85}Rb cell is used to absorb the fraction of the retrieve/EIT control laser not filtered by the polarizing beamsplitters; this requires a retrieve and EIT control laser detuning of 400 MHz. The source and target ensembles are 4.5-cm-long isotopically pure ^{87}Rb vapour cells with 7 torr and 8 torr respectively of neon buffer gas. PBS, polarizing beamsplitter; SMF, single-mode fibre; PM, polarization maintaining; BS, beamsplitter; and S1, S2 (or AS1, AS2) for avalanche photodetectors (APDs) for the Stokes (or anti-Stokes) channel.

high-frequency side of the EIT resonance.

These observations clearly demonstrate that EIT transmission preserves the non-classical statistics of the anti-Stokes pulses. The narrow resonances observed in transmission and photon-correlation data set an upper bound to the bandwidth (of order MHz) of the single-photon pulses generated in our experiments. To analyse these observations, we consider a theoretical model that describes the propagation of single photons of finite bandwidth and purity (that is, a finite probability of two-photon events) in an optically dense, coherently driven medium of three-level atoms. Included in this model is Doppler broadening, realistic detunings (resulting in an asymmetric spectrum), finite decay of the $|g\rangle$ – $|s\rangle$ coherence, and spectrally broad noise associated with two-photon events. As shown in Fig. 3, the theoretical predictions are in good agreement with experimental observations. Note that this analysis shows that the spectral properties of single-photon and two-photon events in conditionally generated pulses differ. These effects, which involve the interplay between spectral and quantum-statistical properties, warrant further investigation.

One intriguing application of EIT involves the controllable delay of optical pulses by slowing their group velocity^{10,11} and stopping their propagation^{12–14,25}. Figure 4 presents an experimental realization of such controllable delay and storage for single-photon pulses. For these measurements, the single-photon anti-Stokes pulses were tuned to the centre of the EIT transmission window ($\delta = 0$); the retrieve laser was turned on for approximately 150 ns, generating anti-Stokes pulses of corresponding duration. Time-resolved measurements shown in Fig. 4a reveal substantial delay, relative to free-space propagation, of the conditionally generated anti-Stokes pulses upon transmission through the EIT medium. As shown in Fig. 4b, we observe delays up to 45 ns in our 4.5-cm-long ensemble, corresponding to single photon velocities of about 10^3 km s^{-1} (~ 0.003 times the speed of light in vacuum). In Fig. 4a, the observed delay of 40 ns corresponds to a substantial fractional delay of about 30% when compared to the 140 ns full-width at half-maximum of the reference pulse.

Figure 4c demonstrates that a fraction of the incoming single-photon pulses can be stored by dynamically reducing the single-photon group velocity to zero. This is accomplished by turning off

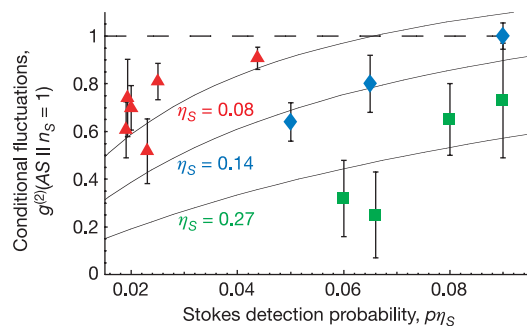


Figure 2 | Observation of conditional single-photon generation. Anti-Stokes fluctuations, conditioned on detection of a single Stokes photon, are characterized by the correlation function $g^{(2)}(\text{AS} || n_s = 1) = \langle \hat{n}_{\text{AS1}} \hat{n}_{\text{AS2}} \rangle / \langle \hat{n}_{\text{AS1}} \rangle \langle \hat{n}_{\text{AS2}} \rangle$, where $(\hat{n}_{\text{AS1}}, \hat{n}_{\text{AS2}})$ is the number operator for detector (AS1, AS2), see Fig. 1b. The dotted line represents the classical limit of $g^{(2)}(\text{AS} || n_s = 1) = 1$. Measurements are shown for three values of the Stokes channel transmission: $\eta_s = 0.08$ (red triangles), $\eta_s = 0.14$ (blue diamonds) and $\eta_s = 0.27$ (green squares). Solid lines represent a theoretical model²⁴ for η_s equal to 0.08, 0.14 and 0.27 respectively. For this data, source ensemble temperature $\sim 26^\circ\text{C}$ (estimated optical depth ~ 4). Anti-Stokes channel transmission is 10%. Experimental repetition rate is 72 kHz. Statistical error bars represent averages of $\sim 400,000$ anti-Stokes detection events, corresponding to total averaging times of ~ 1 hour per point. Error bars, ± 1 s.d.

the EIT control laser as the anti-Stokes pulse propagates in the target ensemble. The stored fraction is released when the control laser is turned back on^{12–14}. Figure 4d shows the conditional storage and retrieval probability as a function of storage time. Storage and retrieval of up to 10% of the incoming pulse was observed at short storage times; retrieved pulses were observed for times up to a few microseconds, limited by atomic diffusion in the target ensemble. Even with these limited efficiencies, the retrieved pulses preserve some non-classical features after considerable storage intervals. For example, for a storage time of $0.5 \mu\text{s}$, we deduce $R = 1.08 \pm 0.01 > 1$. The storage and retrieval efficiency could be improved by, for example, increasing the optical depth or utilizing an optical cavity with modest finesse²⁷. The storage times could be considerably extended by reducing the effect of atomic diffusion, either by expanding the detection-mode diameter, working with ultra-cold atoms in dipole traps or optical lattices, or using a doped solid¹⁴. A factor of ten increase in the detection-mode diameter should extend storage times to a fraction of a millisecond¹².

These results demonstrate that EIT represents a very effective technique for generation and controlled propagation of narrow-bandwidth single-photon light pulses in optically dense atomic

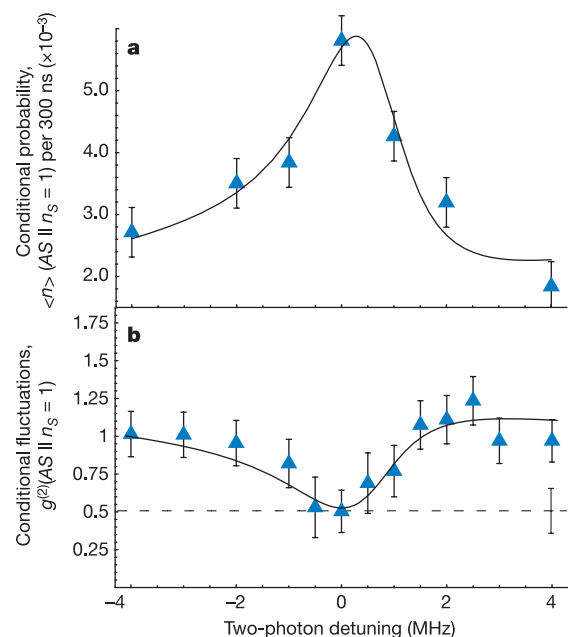


Figure 3 | Observation of single-photon EIT. **a**, Conditional probability (per 300 ns) of detecting an anti-Stokes photon transmitted through the target ensemble, $\langle n \rangle(\text{AS} || n_s = 1)$, versus the EIT two-photon detuning δ . Background (detection probability with write laser off) has been subtracted for transmission data in Figs 3a and 4. For incident pulses, $\langle n \rangle(\text{AS} || n_s = 1) \approx 0.01$. **b**, Second-order correlation function of the anti-Stokes field conditioned on detecting one Stokes photon, $g^{(2)}(\text{AS} || n_s = 1)$, as a function of δ . Dashed line and error bar represent measured value with no target ensemble present. For the data shown, δ is varied by varying the EIT control frequency. For these experiments, $p\eta_s \approx 0.06$, $\eta_s \approx 0.25$ and the (source, target) ensemble temperature $\approx (26^\circ\text{C}, 30^\circ\text{C})$. Overall probability per trial to detect anti-Stokes photons is 6×10^{-3} for incident (target cell absent) photons and 3×10^{-3} for transmitted (target cell present, $\delta = 0$) photons. Statistical error bars represent averages of ~ 1 million anti-Stokes detection events, corresponding to total averaging times of ~ 2 hours per point. The solid line results from the theoretical model described in the text; the parameter values used in the model (retrieve laser Rabi frequency = 35 MHz, retrieve laser detuning = 400 MHz, anti-Stokes background = 0.003 photons per pulse, optical depth = 2.5, single-photon bandwidth = 0.7 MHz, $|g\rangle$ – $|s\rangle$ coherence decay = 0.02 MHz), are similar to experimental conditions. Error bars, ± 1 s.d.

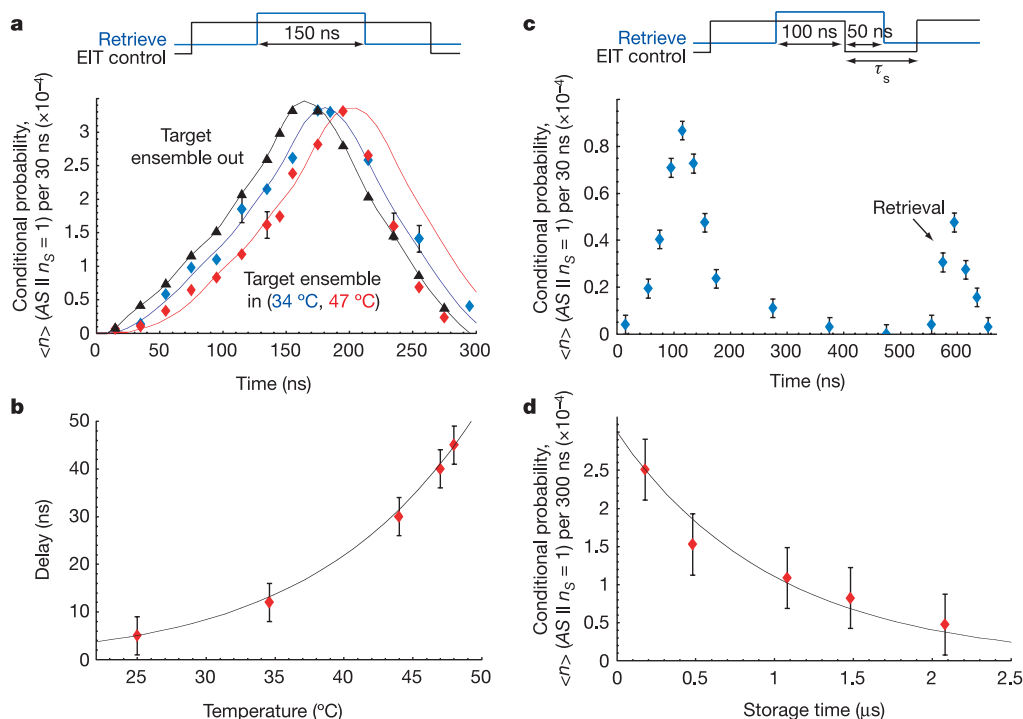


Figure 4 | Time-resolved measurements of single-photon pulse delay and storage. **a**, Conditional probability (per 30 ns) of detecting an anti-Stokes photon transmitted through the target ensemble. Target ensemble absent (black triangles); target ensemble present at 34.6 °C (blue diamonds) and 47 °C (red diamonds). Delayed pulse at (34.6 °C, 47 °C) is scaled by (1.34, 2.14). Solid lines represent theoretical calculations for EIT propagation in a Doppler-broadened medium. **b**, Delay, relative to free-space propagation, of single-photon anti-Stokes pulses, as a function of target-ensemble temperature. Solid line is a theoretical prediction for an EIT control field Rabi frequency of 35 MHz. **c**, Storage and retrieval of a single-photon anti-Stokes pulse. EIT control is turned off 100 ns after the retrieval

from the source ensemble begins; after waiting for a storage time of $\tau_s = 460$ ns, the EIT control is turned back on, resulting in the retrieved pulse centred at 600 ns. Target ensemble temperature ~ 47 °C. **d**, Conditional probability (per 300 ns) of detecting an anti-Stokes photon retrieved from the target ensemble after storage interval τ_s . $\langle n \rangle$ (AS || $n_s = 1$) = 0.003 for the incident pulse (measured with target ensemble absent). For short storage times, the overall probability per trial to detect (incident, retrieved) anti-Stokes photons is $\sim (1.5 \times 10^{-3}, 1.5 \times 10^{-4})$. Decay of probability is fitted by an exponential with a $1/e$ characteristic time of about 1 μ s; this is consistent with diffusion of atoms from the detection volume. Error bars, ± 1 s.d.

ensembles. Applications of quantum-optical processes involving simultaneous control over temporal, spectral, and quantum-statistical properties of single photons are possible^{1,2,28,29}. For example, by storing polarization-encoded qubits either in a pair of atomic ensembles²⁰, or in a pair of different Zeeman sublevels²¹, this technique can be used for exploring quantum-information concepts such as quantum networks¹ and repeaters². At the same time, coherent nonlinear-optical interactions at the single-photon level have been proposed by combining these techniques with resonantly enhanced atomic nonlinearities²⁹.

Received 2 September; accepted 13 October 2005.

- Briegleb, H. J., Dur, W., van Enk, S. J., Cirac, J. I. & Zoller, P. In *The Physics of Quantum Information* (eds Bouwmeester, D., Ekert, A. & Zeilinger, A.) 281–293 (Springer, Berlin, 2000).
- Duan, L. M., Lukin, M. D., Cirac, J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414**, 413–418 (2001).
- Kuzmich, A. *et al.* Generation of nonclassical photon pairs for scalable quantum communication with atomic ensembles. *Nature* **423**, 731–734 (2003).
- van der Wal, C. H. *et al.* Atomic memory for correlated photon states. *Science* **301**, 196–200 (2003).
- McKeever, J. *et al.* Deterministic generation of single photons from one atom trapped in a cavity. *Science* **303**, 1992–1994 (2004).
- Kuhn, A., Hennrich, M. & Rempe, G. Deterministic single-photon source for distributed quantum networking. *Phys. Rev. Lett.* **89**, 067901 (2002).
- Juulsgaard, B., Sherson, J., Cirac, J. I., Fiurasek, J. & Polzik, E. S. Experimental demonstration of quantum memory for light. *Nature* **432**, 482–486 (2004).
- Harris, S. E. Electromagnetically induced transparency. *Phys. Today* **50**, 36–42 (1997).
- Fleischhauer, M., Imamoglu, A. & Marangos, J. P. Electromagnetically induced

- transparency: Optics in coherent media. *Rev. Mod. Phys.* **77**, 633–673 (2005).
- Hau, L. V., Harris, S. E., Dutton, Z. & Behroozi, C. H. Light speed reduction to 17 metres per second in an ultracold atomic gas. *Nature* **397**, 594–598 (1999).
- Kash, M. M. *et al.* Ultraslow group velocity and enhanced nonlinear optical effects in a coherently driven hot atomic gas. *Phys. Rev. Lett.* **82**, 5229–5232 (1999).
- Phillips, D. F., Fleischhauer, A., Mair, A., Walsworth, R. L. & Lukin, M. D. Storage of light in atomic vapor. *Phys. Rev. Lett.* **86**, 783–786 (2001).
- Liu, C., Dutton, Z., Behroozi, C. H. & Hau, L. V. Observation of coherent optical information storage in an atomic medium using halted light pulses. *Nature* **409**, 490–493 (2001).
- Longdell, J. J., Fraval, E., Sellars, M. J. & Manson, N. B. Stopped light with storage times greater than one second using electromagnetically induced transparency in a solid. *Phys. Rev. Lett.* **95**, 063601 (2005).
- Braje, D. A., Balić, V., Yin, G. Y. & Harris, S. E. Low-light-level nonlinear optics with slow light. *Phys. Rev. A* **68**, 041801(R) (2003).
- Chou, C. W., Polyakov, S. V., Kuzmich, A. & Kimble, H. J. Single-photon generation from stored excitation in an atomic ensemble. *Phys. Rev. Lett.* **92**, 213601 (2004).
- Eisaman, M. D. *et al.* Shaping quantum pulses of light via coherent atomic memory. *Phys. Rev. Lett.* **93**, 233602 (2004).
- Jiang, W., Han, C., Xue, P., Duan, L.-M. & Guo, G.-C. Nonclassical photon pairs generated from a room-temperature atomic ensemble. *Phys. Rev. A* **69**, 043819 (2004).
- Balić, V., Braje, D. A., Kolchin, P., Yin, G. Y. & Harris, S. E. Generation of paired photons with controllable waveforms. *Phys. Rev. Lett.* **94**, 183601 (2005).
- Matsukevich, D. N. & Kuzmich, A. Quantum state transfer between matter and light. *Science* **306**, 663–666 (2004).
- Matsukevich, D. N. *et al.* Entanglement of a photon and a collective atomic excitation. *Phys. Rev. Lett.* **95**, 040405 (2005).
- Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics* (Cambridge Univ. Press, Cambridge, 1995).
- Michler, P. *et al.* A quantum dot single-photon turnstile device. *Science* **290**, 2282–2285 (2000).

24. Eisaman, M. D. et al. in *Fluctuations and Noise in Photonics and Quantum Optics III* Vol. 5,842 (eds Hemmer, P. R., Gea-Banacloche, J. R., Heszler, P. & Zubairy, M. S.) 105–113 (SPIE, Bellingham, Washington, 2005).
25. Fleischhauer, M. & Lukin, M. D. Dark-state polaritons in electromagnetically induced transparency. *Phys. Rev. Lett.* **84**, 5094–5097 (2000).
26. Akamatsu, D., Akiba, K. & Kozuma, M. Electromagnetically induced transparency with squeezed vacuum. *Phys. Rev. Lett.* **92**, 203602 (2004).
27. Black, A. T., Thompson, J. K. & Vuletić, V. On-demand superradiant conversion of atomic spin gratings into single photons with high efficiency. *Phys. Rev. Lett.* **95**, 133601 (2005).
28. Scully, M. O. & Ooi, C. H. R. Improving quantum microscopy and lithography via Raman photon pairs: II. *Analysis. J. Opt. B* **6**, S816–S820 (2004).
29. André, A., Bajcsy, M., Zibrov, A. S. & Lukin, M. D. Nonlinear optics with stationary pulses of light. *Phys. Rev. Lett.* **94**, 063902 (2005).
30. André, A. *Nonclassical States of Light and Atomic Ensembles: Generation and New Applications* PhD thesis, Harvard Univ. (2005).

Acknowledgements We acknowledge T. Zibrova, A. Gorshkov, P. Hemmer, J. MacArthur, D. Phillips and R. Walsworth for discussions and experimental help. This work was supported by DARPA, the Packard and Sloan Foundations, and the NSF through the CAREER programme and the Harvard-MIT Center for Ultracold Atoms.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.D.E. (eisaman@fas.harvard.edu).

LETTERS

A lithospheric instability origin for Columbia River flood basalts and Wallowa Mountains uplift in northeast Oregon

T. C. Hales¹, D. L. Abt^{1†}, E. D. Humphreys¹ & J. J. Roering¹

Flood basalts appear to form during the initiation of hotspot magmatism. The Columbia River basalts (CRB) represent the largest volume of flood basalts associated with the Yellowstone hotspot, yet their source appears to be in the vicinity of the Wallowa Mountains¹, about 500 km north of the projected hotspot track. These mountains are composed of a large granitic pluton intruded into a region of oceanic lithosphere affinity². The elevation of the interface between Columbia River basalts and other geological formations indicates that mild pre-eruptive subsidence took place in the Wallowa Mountains, followed by syn-eruptive uplift of several hundred metres and a long-term uplift of about 2 km. The mapped surface uplift mimics regional topography, with the Wallowa Mountains in the centre of a 'bull's eye' pattern of valleys and low-elevation mountains. Here we present the seismic velocity structure of the mantle underlying this region and erosion-corrected elevation maps of lava flows, and show that an area of reduced mantle melt content coincides with the 200-km-wide topographic uplift. We conclude that convective downwelling and detachment of a compositionally dense plutonic root can explain the timing and magnitude of Columbia River basalt magmatism, as well as the surface uplift and existence of the observed melt-depleted mantle.

The Yellowstone hotspot initiated ~17 Myr ago with a rapid succession of large eruptions starting at the McDermitt caldera in northern Nevada and propagating ~500 km north along the western edge of the Precambrian continental margin, first to Steens Mountains area in southeast Oregon and then to the site of the CRB eruptions in northeast Oregon¹ (Fig. 1). The total volume of erupted flood basalt is ~234,000 km³, 75% of which is CRB (ref. 3). The large geographic extent involved in these eruptions has been difficult to reconcile with a traditional plume model because nearly all flood basalt erupted north of the projected hotspot track, which trends from Yellowstone to the McDermitt caldera. Furthermore, the Snake River plain is a well-developed hotspot track, but it leads away from the CRB source area and arcs across the state of Idaho like a smile (Fig. 1 inset). Two recent suggestions accounting for the distributed and largely off-track nature of hotspot initiation modify the plume hypothesis by having plume impingement beneath either the Steens Mountains¹ or the central Idaho craton⁴ and then flattening rapidly along a trend near the Precambrian continental margin. These plume models are supported by the presence of high ³He/⁴He in the flood basalts⁵. In contrast, other models suggest that flood basalt magmatism represents passive back-arc processes⁶.

At present, little deformation occurs in northeast Oregon. Global Positioning System (GPS) and geologic data indicate clockwise rotation of much of Oregon relative to North America about a

pole near the bend in the Precambrian continental margin (Fig. 1), consistent with regional tectonics^{7,8}. E–W extension on normal faults occurs with increasing rate south of this location, and N–S contraction, expressed as E–W-oriented folds, is increasingly active to the west. This regional tectonic setting has persisted since eruption of the CRB without noticeable change^{7,9}, and proximity to the pole implies very low deformation rates for northeast Oregon. The regional ~N–S axis of maximum compression^{9–12} responsible for these faults and folds was also responsible for the NNW–SSE trending Chief Joseph dike swarm, E–W folds such as the Lewiston anticline and NW–SE-oriented strike-slip structures such as the Olympic–Wallowa Lineament^{10–12}. Some young structures near the Wallowa Mountains have deformed in an exceptional manner that is inconsistent with the regional stress field. These include the WNW–ESE-oriented Wallowa normal fault, the oblique normal Limekiln and Hite Fault systems and the NE–SW-oriented folds such as the Troy basin syncline¹². We suggest that these inconsistencies with the regional stress field represent local perturbations that may be caused by local contrasts in strength or buoyancy.

CRB lavas erupted onto a low-relief surface of Mesozoic accreted ocean terranes stitched together by Jurassic plutons, of which the Wallowa pluton is the largest¹³ (Fig. 1). After initial Imnaha flows filled the valleys, subsequent eruptions deposited thin, flat-lying sheets that provide a useful datum to measure deformation accurately. These flows are found in Hells Canyon at an elevation of 1,100 m and atop the Wallowa Mountains at 2,700 m, demonstrating substantial vertical motion following their eruption¹⁴. Grande Ronde eruptions followed the Imnaha, depositing about 80% of the CRB lavas over ~1 Myr. We correct elevation maps of the Grande Ronde magnetostratigraphic units (which are subdivided into R₁, N₁, R₂ and N₂ units)^{14–16} for the effects of erosional unloading^{17,18} to obtain markers of post-eruptive deformation (Fig. 2, Supplementary Figure S1). The uplift and downwarp of these flows closely conform to current topography, with up to 2 km of uplift focused on the Wallowa pluton (Figs 1 and 2) and more subdued uplift centred on the lesser plutons in the area (Fig. 1). The history of Imnaha and Grande Ronde uplift reveals the early evolution of local buoyancy. Imnaha basalts ponded in incipient basins, indicating possible pre-eruptive subsidence¹⁹. Basin development continued during Grande Ronde eruptions, causing local ponding and thickening of flows, while syn-eruptive uplift created locally thinned units. In particular, ~300 m of uplift resulted in a thinned N₂ unit near its source close to the Wallowa Mountains^{19,20} (Supplementary Fig. S1). Post-Grande-Ronde deformation is evidenced by the contrasting distributions of Grande Ronde and late CRB Wanapum (15–14 Myr ago) and Saddle Mountains (14–6 Myr ago) flows, which show distinctive,

¹Department of Geological Sciences, 1272 University of Oregon, Eugene, Oregon 97403, USA. [†]Present address: Department of Geological Sciences, Brown University, 324 Brook Street, Providence, Rhode Island 02912, USA.

well-developed channelization and ponding in response to emerging topographic relief²¹. Significant relief generation occurred at 14.5–13 Myr ago, coincident with initiation of the La Grande, Baker and Weiser grabens, which surround the Wallowa Mountains¹⁹. This structural evidence leads us to conclude that most Wallowa Mountains uplift occurred in <10 Myr (Fig. 2c).

The current physical state of the upper mantle provides a constraint on the origin of buoyancy. We obtain a basic assessment of upper mantle structure by tomographically inverting ~600 teleseismic P-wave delays collected by a six-station array of IRIS-PASSCAL seismometers deployed in the region. Relatively high-velocity mantle is imaged beneath the bull's eye region (Fig. 3). Resolution 'squeezing' tests²² (Fig. 3b) indicate that this seismically fast volume extends from ~75 km to at most a depth of 175 km, and its horizontal extent is coincident with the bull's eye pattern (Fig. 3). The occurrence of basaltic volcanism in the recent past and the major uplift in the Wallowa Mountains are inconsistent with the high-velocity mantle being relatively cool. The only reasonable alternative is that melt content has been reduced (and possibly eliminated) by eruption of the CRB from the high-velocity volume^{22,23}. This hypothesis is strengthened by the spatial correlation between the upper mantle high-velocity volume and the location of flood basalt dykes (Figs 1 and 3). The implied mantle depletion would increase mantle buoyancy by an amount that accounts for the observed several hundred metres of syn-eruptive uplift²⁴. Using the equations of Schutt and Leshner²⁴, a simple calculation suggests that 5% depletion of an 80-km-thick layer (for example, 70–150 km) would

cause ~300 m of isostatic uplift in the region. However, mantle basalt depletion cannot account for the focused ~2 km of surface uplift of the Wallowa Mountains.

The unusual tectonic and magmatic history of northeast Oregon has similarities to two recently studied events that involve apparent small-scale convective instability driven by dense garnet-rich crust and mantle lithosphere. Elkins-Tanton and Hager²⁵ account for pre-eruptive subsidence of the Siberian traps flood basalts with the initial downwelling of dense eclogitic lower lithosphere beneath the area of downwarping, and attribute the magmatism to decompression melting of ascending asthenosphere flowing into the vacated regions. Saleeby and Foster²⁶ compile abundant evidence for young and ongoing convective descent (termed 'delamination' in their paper) of the garnet-rich rocks that are the genetic complement to the granitic Sierra Nevada pluton in California. As this dense lower crust and upper mantle sinks, the overlying southern Sierra Nevada Mountains rise in isostatic response.

In northeast Oregon, where granitic plutons are scarce, major uplift is concentrated on the largest pluton in the region. We attribute this uplift to the convective removal of the pluton's dense roots, and consider the magnitude of uplift and its strong association with the Wallowa pluton difficult to account for with any other process. Furthermore, such a delamination-driven model can also explain the remarkable bull's eye uplift pattern, the post-eruptive timing of most uplift, and the associated CRB outpourings. We envision initial convective instability of the plutonic roots creating a drip-like downwelling that pulls the Wallowa region down, creating

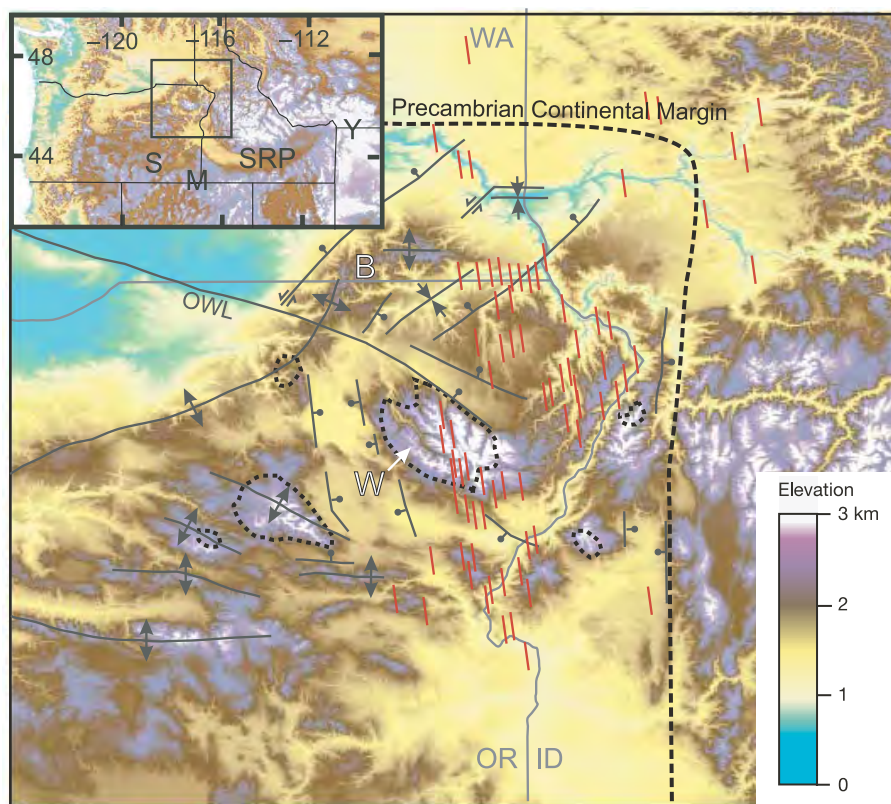


Figure 1 | Regional structural and location maps for northeast Oregon.

Location map showing CRB source area, the Wallowa (W) and Blue (B) Mountains. Productive Grande Ronde dike swarms are shown as short red lines drawn in the average orientation of these features. These dike swarms show the approximate location of major dike swarms and the line length does not represent the actual length of dike swarms. Major extensional (lines with balls) and contractional (lines with double-headed arrows) structures are indicated¹². The Precambrian continental margin, represented with $^{87}\text{Sr}/^{86}\text{Sr} = 0.706$ in areas where it is not exposed, is indicated with a dashed line. Major plutons

west of Precambrian North America are outlined with heavy dotted lines. WA, Washington; OR, Oregon; ID, Idaho; OWL, Olympic-Wallowa Lineament. Inset, map of the western United States showing locations of major magmatic centres in relation to the Yellowstone hotspot, including: the McDermitt caldera (M), the Steens Mountains (S) and the source of the CRB eruptions in northeast Oregon (small rectangle). The arcuate Snake River Plain (SRP) is the Yellowstone hotspot track leading to the active Yellowstone caldera (Y). US State boundaries are shown as solid lines. Note the bull's eye pattern of uplift at the northwest end of the SRP.

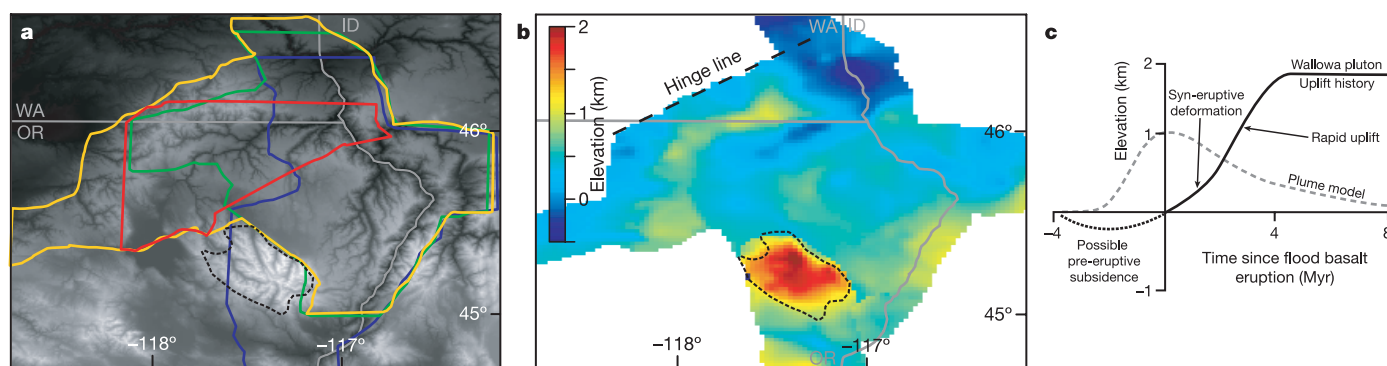


Figure 2 | Maps of post-eruptive uplift in the Wallowa Mountains area. **a**, Digital elevation map showing the overlapping distribution of exposed Grande Ronde magmatostratigraphic interfaces used for analysis of Wallowa Mountains uplift: Imnaha-R1 (blue), R1-N1 (green), N1-R2 (yellow) and R2-N2 (red). **b**, Composite surface for Grande Ronde uplift. This surface was made by vertically shifting the individual interfaces using published flow thicknesses¹⁴, smoothing the resulting surface with a mild low-pass filter, and removing the effects of erosional unloading by deconvolving the point load response of an elastic plate^{17,18}, assuming total coverage by the Grande Ronde lavas and an elastic thickness of 5 km, similar to other estimates from this area²². The absolute elevation is referenced to a hinge line (dotted black line) that separates the uplifted Blue Mountains from the down-dropped Pasco basin immediately northwest of the Blue Mountains. **c**, Schematic plot

of the Wallowa Mountains uplift history (dark line) and a hypothetical area affected by a thermal plume (grey line). Wallowa Mountain uplift history is controlled at three times. Pre-eruptive subsidence, inferred from ponding of early Imnaha lava flows in basins, is interpreted as the initiation of delamination. We calculated ~300 m of syn-eruptive uplift in this area over the 1-Myr duration of the Grande Ronde eruption. This is followed by rapid uplift of the Wallowa Mountains 3–4 Myr after initial Imnaha eruption, indicated by the development of structures bounding the Wallowa Mountains¹⁴, and current topographic relief of the composite surface shown in **b**. We note that the hypothetical uplift curve for a thermal plume can vary significantly in magnitude, so we have chosen a representative amount of uplift predicted by CRB magmatism.

pre-eruptive subsidence (Fig. 2c). Asthenosphere flowing up into the area around the Wallowa downwelling decompresses and partly melts upon ascent, creating initial CRB flows that pond in the depression. As the dense mantle and lower crust sink and detach from the lithosphere, surface uplift begins and magmatism accelerates within the circular area where decompression melting is excited. The uplift of ~300 m that occurred during this time interval is controlled by the combined effects of melt accumulation in magma chambers and chamber evacuation, and residual buoyancy created by basalt depletion. Finally, continued sinking of the dense Wallowa pluton root causes continued uplift of the Wallowa pluton, which progresses until the sinking mass is effectively

removed from the region and uplift becomes fully isostatic (Fig. 2c).

This model for flood basalt volcanism driven by delamination of compositionally dense roots has some interesting implications for Yellowstone hotspot initiation. We discuss two interpretations: First, if a plume impacted to the south of the CRB source area and propagated north along the Precambrian continental margin^{1,4}, associated hot mantle material could trigger convective lithospheric instability of the Wallowa plutonic roots. This mechanism could thus provide a way of accounting for the intense magmatism far from the Yellowstone hotspot track. In this case, a significant modification to the plume hypothesis is not required and the problem of CRB magmatism is addressed. Second, the sequence of eruptions initiating the Yellowstone hotspot could have been caused by a series of delamination events, perhaps one triggering the next or all triggered by unusual back-arc activity, with no requirement for an active plume. In this case, magmatism could be related to a swath of fertile mantle lithosphere adjacent to the infertile Precambrian lithosphere that was recently hydrated in a fore-arc setting. In either case, delamination is required to initiate flood basalt volcanism in northeast Oregon, and we have shown there to be an intimate relationship between crustal structure and flood basalt initiation.

Received 14 March; accepted 6 October 2005.

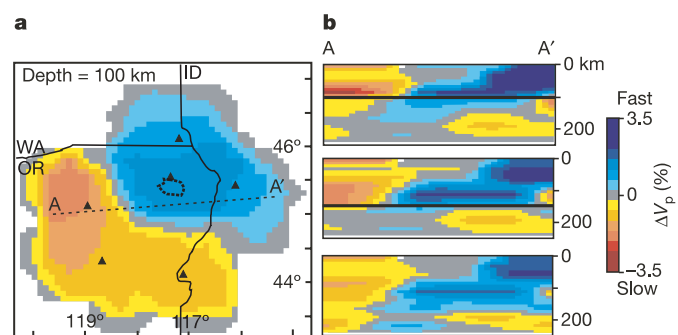


Figure 3 | Tomographic images of northeast Oregon. **a**, Upper mantle P-wave velocity variations at 100 km depth, imaged by teleseismic tomography from data recorded by six seismometers (triangles). Also shown is the Wallowa pluton (dotted line) and cross-section location (dashed line). **b**, Cross-sections through 'squeezed' inversions. These are hypothesis tests in which we found the structure that best accounts for the observed delays and that is confined above the black line; after relaxing the depth constraint, creation of significant structure below the black line indicates that deeper structure is required. The upper cross-section (squeezing depth = 100 km) indicates a strong need for structure below 100 km. The middle panel (squeezing depth = 150 km, and used for **a**), like the lower (unconstrained) panel, indicates no need for high-velocity structure below a depth of 150 km. Imaging places the top of the high-velocity body near 75 km. ΔV_p represents the percentage difference from average P-wave velocity.

1. Camp, V. E. & Ross, M. E. Mantle dynamics and genesis of mafic magmatism in the intermontane Pacific Northwest. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002838 (2004).
2. Lallemand, H. G. A. Pre-Cretaceous tectonic evolution of the Blue Mountains province, northeastern Oregon. *USGS Prof. Pap.* **1438**, 271–304 (1994).
3. Camp, V. E., Ross, M. E. & Hanson, W. E. Genesis of flood basalts and basin and range volcanic rocks from Steens Mountain to the Malheur River Gorge, Oregon. *Geol. Soc. Am. Bull.* **115**, 105–128 (2003).
4. Jordan, B. T. in *Plumes, Plates and Paradigms* (eds Foulger, G. R., Natland, J. H., Presnall, D. C. & Anderson, D. L.) 503–515 (Geological Society of America Memoir 388, Boulder, Colorado, 2005).
5. Dodson, A., Kennedy, M. B. & De Paolo, D. J. Helium and neon isotopes in the Imnaha Basalt, CRBG, evidence for a Yellowstone plume source. *Earth Planet. Sci. Lett.* **150**, 443–451 (1997).
6. Anderson, D. L. The sublithospheric mantle as the source of continental flood basalts; the case against the continental lithosphere and plume head reservoirs. *Earth Planet. Sci. Lett.* **123**, 269–280 (1994).

7. Wells, R. E., Weaver, C. S. & Blakely, R. J. Fore-arc migration in Cascadia and its neotectonic significance. *Geology* **26**, 759–762 (1998).
8. McCaffrey, R. *et al.* Rotation and plate locking at the southern Cascadia subduction zone. *Geophys. Res. Lett.* **27**, 3117–3120 (2000).
9. Hooper, P. R. The Columbia River flood basalt province: current status. *Geophys. Monogr.* **100**, 1–27 (1997).
10. Hooper, P. R. & Camp, V. E. Deformation of the southeast part of the Columbia plateau. *Geology* **9**, 323–328 (1981).
11. Camp, V. E. & Hooper, P. R. Geologic studies of the Columbia Plateau. Part 1. Late Cenozoic evolution of the southeastern part of the Columbia River basalt province. *Geol. Soc. Am. Bull.* **92**, 659–668 (1981).
12. Hooper, P. R. & Conrey, R. M. A model for the tectonic setting of the Columbia River basalt eruptions. *Geol. Soc. Am. (Spec. Pap.)* **239**, 293–306 (1989).
13. Goles, G. G. Miocene basalts of the Blue Mountains province in Oregon. I. Compositional types and their geological settings. *J. Petrol.* **27**, 495–520 (1986).
14. Walker, G. W. *Reconnaissance Geologic Map of the Oregon Part of the Grangeville Quadrangle, Baker, Union, Umatilla, and Wallowa Counties, Oregon Map* (USGS Misc. Investigations Series I-1116, 1979).
15. Reidel, S. P., Hooper, P. R., Webster, G. D. & Camp, V. E. *Geologic Map of Southeastern Asotin County, Washington* (Washington Division of Geology and Earth Resources Map Series, GM-40, 1992).
16. Hooper, P. R., Webster, G. D. & Camp, V. E. *Geologic Map of the Clarkson 15 Minute Quadrangle, Idaho and Washington* (Washington Division of Geology and Earth Resources Map Serie GM-31, 1985).
17. Lambeck, K. *Geophysical Geodesy: the Slow Deformations of the Earth* Ch. 10 (Clarendon, Oxford, 1988).
18. Anderson, R. S. Evolution of the Santa Cruz Mountains, California, through tectonic growth and geomorphic decay. *J. Geophys. Res.* **99**, 20161–20179 (1994).
19. Camp, V. E. Geologic studies of the Columbia Plateau: Part II. Upper Miocene basalt distribution, reflecting source locations, tectonism, and drainage history in the Clearwater embayment Idaho. *Geol. Soc. Am. Bull.* **92**, 669–678 (1981).
20. Reidel, S. P. *et al.* The Grande Ronde Basalt, Columbia River Basalt Group; Stratigraphic descriptions and correlations in Washington, Oregon, and Idaho. *Geol. Soc. Am. (Spec. Pap.)* **239**, 21–54 (1989).
21. Tolan, T. L. *et al.* Revisions to the estimates of the areal extent and volume of the Columbia River Basalt Group. *Geol. Soc. Am. (Spec. Pap.)* **239**, 1–20 (1989).
22. Saltzer, R. L. & Humphreys, E. D. Upper mantle P-wave structure of the eastern Snake River Plain and its relationship to geodynamic models of the region. *J. Geophys. Res.* **102**, 11829–11841 (1997).
23. Hammond, W. C. & Humphreys, E. D. Upper mantle seismic wave velocity: The effect of realistic partial melt geometries. *J. Geophys. Res.* **105**, 10975–10999 (2000).
24. Schutt, D. L. & Leshner, C. E. The effect and seismic velocity of garnet and spinel lherzolite. *J. Geophys. Res.* (in the press).
25. Elkins-Tanton, L. T. & Hager, B. H. Melt intrusion as a trigger for lithospheric foundering and the eruption of the Siberian flood basalts. *Geophys. Res. Lett.* **27**, 3937–3940 (2000).
26. Saleeby, J. B. & Foster, Z. Topographic response to mantle lithosphere removal in the southern Sierra Nevada region, California. *Geology* **32**, 245–248 (2004).
27. Armstrong, R. L., Taubeneck, W. H. & Hales, P. O. Rb-Sr and K-Ar geochronometry of Mesozoic granitic rocks and their Sr isotopic composition, Oregon, Washington and Idaho. *Geol. Soc. Am. Bull.* **88**, 387–411 (1977).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Goles for initial conversations regarding this manuscript. Funding for this work provided by National Science Foundation grants to J.J.R. and E.D.H. A review from M. Richards improved the manuscript. We thank J. Riker, D. Clippinger, C. Simmons and M. Giba for field assistance. We also thank N. Fay and J. Crosswhite for help with numerical analysis.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.H. (thales@uoregon.edu).

LETTERS

Determinants of woody cover in African savannas

Mahesh Sankaran¹, Niall P. Hanan¹, Robert J. Scholes², Jayashree Ratnam¹, David J. Augustine³, Brian S. Cade⁴, Jacques Gignoux⁵, Steven I. Higgins⁶, Xavier Le Roux⁷, Fulco Ludwig⁸, Jonas Ardo⁹, Feetham Banyikwa¹⁰, Andries Bronn¹¹, Gabriela Bucini¹, Kelly K. Caylor¹², Michael B. Coughenour¹, Alioune Diouf¹³, Wellington Ekaya¹⁴, Christie J. Feral¹⁵, Edmund C. February¹⁶, Peter G. H. Frost¹⁷, Pierre Hiernaux¹⁸, Halszka Hrabar¹⁹, Kristine L. Metzger²⁰, Herbert H. T. Prins²¹, Susan Ringrose²², William Sea¹, Jörg Tews²³, Jeff Worden¹ & Nick Zambatis²⁴

Savannas are globally important ecosystems of great significance to human economies. In these biomes, which are characterized by the co-dominance of trees and grasses, woody cover is a chief determinant of ecosystem properties^{1–3}. The availability of resources (water, nutrients) and disturbance regimes (fire, herbivory) are thought to be important in regulating woody cover^{1,2,4,5}, but perceptions differ on which of these are the primary drivers of savanna structure. Here we show, using data from 854 sites across Africa, that maximum woody cover in savannas receiving a mean annual precipitation (MAP) of less than ~650 mm is constrained by, and increases linearly with, MAP. These arid and semi-arid savannas may be considered ‘stable’ systems in which water constrains woody cover and permits grasses to coexist, while fire, herbivory and soil properties interact to reduce woody cover below the MAP-controlled upper bound. Above a MAP of ~650 mm, savannas are ‘unstable’ systems in which MAP is sufficient for woody canopy closure, and disturbances (fire, herbivory) are required for the coexistence of trees and grass. These results provide insights into the nature of African savannas and suggest that future changes in precipitation⁶ may considerably affect their distribution and dynamics.

Savannas occupy a fifth of the earth's land surface and support a large proportion of the world's human population and most of its rangeland, livestock and wild herbivore biomass¹. A defining feature of savanna ecosystems is the coexistence of trees and grasses in the landscape¹. The balance between these two life forms influences both plant and livestock production, and has profound impacts on several aspects of ecosystem function, including carbon, nutrient and hydrological cycles^{1–3,7}. The mechanisms that promote tree–grass coexistence and the factors that determine the relative proportions of these two life forms across different savanna types remain, however, unclear^{1,2,4,5}. Because savannas are anticipated to be among the ecosystems that are most sensitive to future changes in land use and climate^{8–10}, a thorough understanding of factors that structure

savanna communities is urgently required to guide management efforts^{2,4}.

Explanations for the persistence of tree–grass mixtures in savannas are varied and invoke such different mechanisms as competition for water and nutrients^{11–13}, demographic bottlenecks to tree recruitment^{5,14}, and disturbances including fire^{1,5,14–17} and large mammal herbivory^{15,16,18}. Empirical studies provide support both for and against each alternative mechanism and, consequently, perceptions differ on the relative importance of resource limitation versus disturbances in controlling savanna structure^{1,2,4,5}. The lack of consensus arises, in part, because most studies have been small-scale and site-specific, and have often focused on a single determinant². But savanna systems are diverse and occur under a wide range of bioclimatic conditions², and it is likely that the importance of different processes in regulating woody cover may vary in different savanna regions. Thus, a comprehensive model that explains both coexistence and the relative productivity of tree and grass components across diverse types of savanna is unlikely to arise from studying individual systems in isolation: it requires a synthesis of data from savannas across broad environmental gradients^{2,4}.

Here we use a continental scale analysis of African savannas to investigate how the relative importance of resource availability (water, nutrients) and disturbance regimes (fire, herbivory) in regulating woody cover varies across broad environmental gradients. In particular, we are interested in determining whether broad-scale trends in savanna structure are indicative of ‘stable’ or ‘unstable’ dynamics¹, or whether savannas show elements of both across their geographic range of occurrence. We use ‘stable’ in a limited sense to mean that coexistence of trees and grasses in savannas is not dependent on disturbances such as fire and mammalian herbivory, while recognizing that woody community biomass and cover are dynamic, not static, properties of the system.

Specifically, we considered that if water availability is the primary determinant of woody cover in savannas^{11–13}, then precipitation

¹Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, Colorado 80523, USA. ²Division of Forest Science and Technology, CSIR, PO Box 395, Pretoria 001, South Africa. ³USDA Forest Service, Commanche National Grassland, PO Box 12, Springfield, Colorado 81073, USA. ⁴US Geological Survey, Fort Collins Science Center, 2150 Centre Avenue, Building C, Fort Collins, Colorado 80526-8818, USA. ⁵Ecole Normale Supérieure, Laboratoire d'écologie, UMR 7625 CNRS, Université de Paris 6 ENS, 46 Rue d'Ulm, 75230 Paris Cedex 05, France. ⁶Umweltforschungszentrum Leipzig Halle, Sekt Ökosyst Anal, Permoserstrasse 15, D-04318 Leipzig, Germany. ⁷Microbial Ecology Laboratory, UMR 5557 CNRS, Université Lyon 1 USC INRA 1193, Bâtiment G. Mendel, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne, France. ⁸CSIRO Centre for Environment and Life Sciences, CSIRO Plant Industry, Private Bag 5, Wembley, Western Australia 6913, Australia. ⁹Department of Physical Geography and Ecosystems Analysis, Lund University, Sölvegatan 12, 223 62 Lund, Sweden. ¹⁰Department of Botany, University of Dar es Salaam, PO Box 35060, Dar es Salaam, Tanzania. ¹¹Department of Agriculture and Game Management, Private Bag X6011, Port Elizabeth Technikon, Port Elizabeth 6000, South Africa. ¹²Department of Civil & Environmental Engineering, Princeton University, Princeton, New Jersey 08544, USA. ¹³Centre de Suivi Ecologique, BP 15532, Dakar, Senegal. ¹⁴University of Nairobi, Department of Range Management, PO Box 29053, Nairobi, Kenya. ¹⁵Environmental Sciences Department, University of Virginia, PO Box 400123, 291 McCormick Road, Charlottesville, Virginia 22904-4123, USA. ¹⁶Department of Botany, University of Cape Town, University Private Bag, Rondebosch 7700, South Africa. ¹⁷Institute of Environmental Studies, University of Zimbabwe, PO Box MP 167, Mount Pleasant, Harare, Zimbabwe. ¹⁸CESBIO, 18 Avenue E. Belin, 31401 Toulouse Cedex 9, France. ¹⁹Mammal Research Institute, University of Pretoria, Pretoria 002, South Africa. ²⁰Department of Zoology, University of Wisconsin, 430 Lincoln Drive, Madison, Wisconsin 53706, USA. ²¹Resource Ecology Group, Wageningen University, Bornsesteeg 69, 6708 PD Wageningen, The Netherlands. ²²Harry Oppenheimer Okavango Research Centre, University of Botswana, Private Bag 285, Maun, Botswana. ²³University of Potsdam, Institute of Biochemistry & Biology, Plant Ecology & Nature Conservation, Maulbeerallee 2, D-14469 Potsdam, Germany. ²⁴Scientific Services, Kruger National Park, Private Bag X402, Skukuza 1350, South Africa.

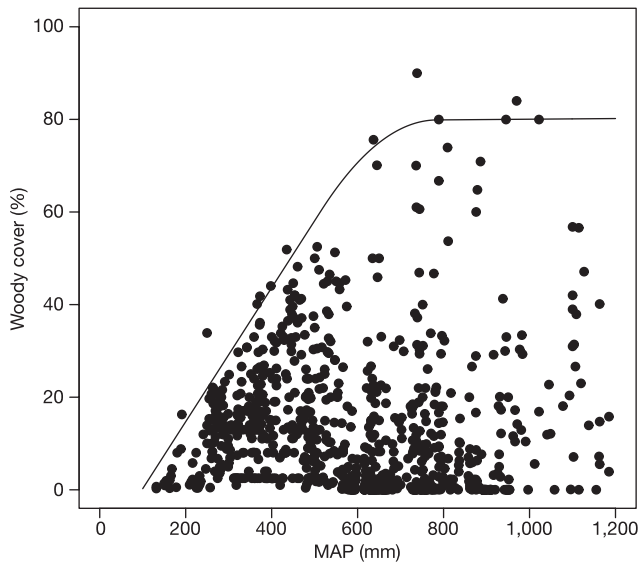


Figure 1 | Change in woody cover of African savannas as a function of MAP. Maximum tree cover is represented by using a 99th quantile piecewise linear regression. The regression analysis identifies the breakpoint (the rainfall at which maximum tree cover is attained) in the interval 650 ± 134 mm MAP (between 516 and 784 mm; see Methods). Trees are typically absent below 101 mm MAP. The equation for the line quantifying the upper bound on tree cover between 101 and 650 mm MAP is $\text{Cover}(\%) = 0.14(\text{MAP}) - 14.2$. Data are from 854 sites across Africa.

should limit the potential tree cover that can be supported at any given site, and maximum realizable woody cover should gradually increase with MAP^{4,12}. By contrast, if disturbances such as fire and herbivory primarily maintain savannas^{4,5,15}, then we expect an abrupt, rather than gradual, increase in maximum realizable woody cover with increasing MAP⁴: below a critical threshold of rainfall sufficient to permit tree growth outside riparian areas or depressions, grasslands should dominate; above this threshold, the maximum woody cover should correspond to a closed-canopy woodland state⁴. Depending on the level of disturbance, a particular location might have reduced woody cover, but the upper bound would not depend on MAP.

We evaluated relationships between woody cover and MAP, soil characteristics (texture, percentage nitrogen, nitrogen mineralization, total phosphorus) and disturbance regimes (fire-return intervals, mammalian herbivore biomass) from 854 sites across Africa (Supplementary Fig. S1 and Methods). Woody cover ranges from 0 to 90% across sites and tends to increase with MAP (Fig. 1). More particularly, within a narrow range of MAP from ~100 to 650 mm, an upper bound exists on the maximum realizable woody cover (Fig. 1). In these arid to semi-arid sites ($<650 \pm 134$ mm MAP; see Fig. 1), maximum realized woody cover increases with MAP (Fig. 2a), but shows no relationship with fire-return intervals, herbivore biomass or soil characteristics (Fig. 2b–f), suggesting that the observed upper limit on woody cover in arid and semi-arid African savannas is primarily a consequence of moisture limitation. The presence of an upper bound on woody cover in these savannas that is linked primarily to MAP is not consistent with the view that savannas are inherently unstable systems maintained by disturbances.

Within this MAP range ($<650 \pm 134$ mm MAP), our analysis suggests that tree–grass coexistence is stable to the extent that disturbances such as fire and herbivory, although capable of modifying tree to grass ratios, are not necessary for coexistence. In these “climatically determined savannas”¹⁷ ($<650 \pm 134$ mm MAP), restrictions on maximum woody cover as a result of water limitation permit grasses to persist in the system. By contrast, in areas that

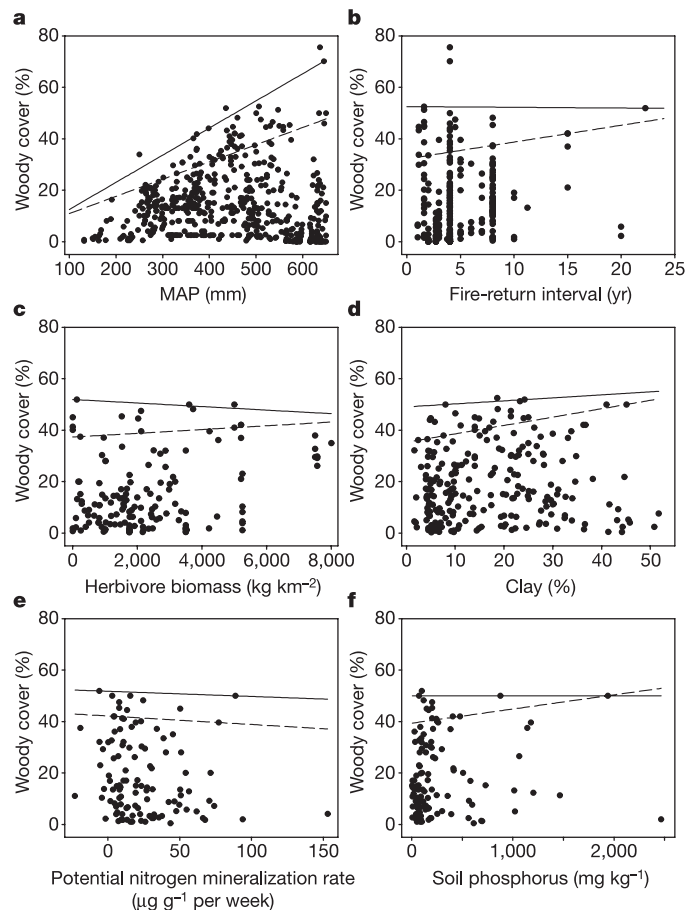


Figure 2 | Woody cover as a function of MAP, soil properties and disturbance regimes in arid and semi-arid savannas. Relationships between woody cover and MAP (a; $n = 529$), fire-return intervals (b; $n = 302$), herbivore biomass (c; $n = 145$), percentage of clay (d; $n = 234$), nitrogen mineralization potential (e; $n = 109$) and soil total phosphorus (f; $n = 118$) for savannas receiving <650 mm MAP. Unbroken and broken lines represent the 99th and 90th linear quantiles, respectively. Maximum woody cover increased with MAP, but showed no consistent relationship with other variables. For MAP, both quantile slopes were significantly different from zero. For fire-return intervals, herbivore biomass, clay and nitrogen mineralization rates, neither regression line had a significant non-zero slope. For total phosphorus, the 90th but not the 99th quantile slope differed from zero.

receive a MAP in excess of 650 ± 134 mm, water availability seems to be sufficient to allow trees to approach canopy closure such that grasses may be effectively excluded. These “disturbance-driven savannas”¹⁷ represent unstable systems in which disturbances such as fire, grazing and browsing are required to maintain both trees and grasses in the system by buffering against transitions to a closed-canopy state^{5,17}.

Whereas MAP drives the upper bound on woody cover in arid and semi-arid savannas, disturbance regimes and soil characteristics impose significant controls on savanna structure by influencing woody cover below the bound. A regression tree analysis of mean woody cover for a restricted subset of sites for which all data were available (Fig. 3 and Methods) further highlights the importance of MAP as a principal driver of savanna structure and suggests that MAP also mediates the relative importance of other savanna drivers such as fire and soil characteristics.

Below a MAP of ~350 mm, woody cover is typically low (Fig. 3). In these sites, soil properties and disturbances such as fire and herbivory rarely regulate woody cover. As MAP increases above this threshold, fire in particular becomes a common factor that reduces woody cover

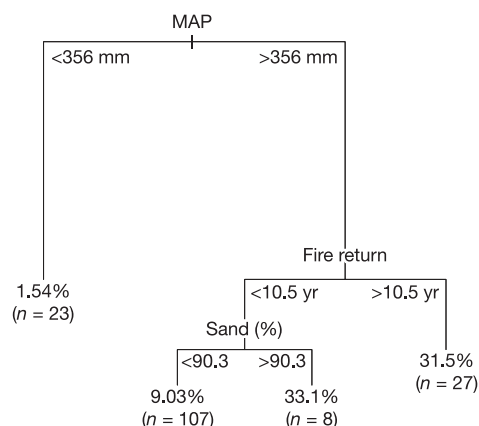


Figure 3 | Regression tree showing generalized relationships between woody cover and MAP, fire-return interval and percentage of sand. The tree is pruned to four terminal nodes and is based on 161 sites for which all data were available. No consistent herbivore effects were detected. Branches are labelled with criteria used to segregate data. Values in terminal nodes represent mean woody cover of sites grouped within the cluster. The pruned tree explained ~45.2% of the variance in woody cover, which is significantly more than a random tree ($P < 0.001$). Of this, 31% was accounted for by the first split; the second split explained an additional 10% of the variance in woody cover.

below the MAP-controlled upper bound (Fig. 3). Woody cover is higher, on average, where fires are infrequent (fire-return interval >10.5 yr). In sites with more frequent fires, woody cover is typically low, except on very sandy soils (mostly concentrated on the Kalahari sand sheets), which tend to support higher woody cover (Fig. 3). The dependence of fire frequency on MAP presumably arises because increased grass production in mesic sites leads to greater fuel loads that can support more frequent fires¹⁴ (Supplementary Fig. S2). Very high sand content, which correlates with low nutrient availability (Supplementary Table S1), may promote higher woody cover if the positive effects of coarse-textured soils, such as lower wilting points¹⁹ and greater water percolation to soil layers below grass rooting depths^{1,11,12}, override the negative effects associated with lower nutrient availability in these soils¹⁹.

Herbivore effects on woody cover are, however, less apparent. Although we found a tendency for grazers to enhance woody cover and browsers and mixed feeders to depress it, such effects were weak and could not be generalized beyond our data set (see Methods; measures of herbivore biomass were retained in the complete, but not pruned, regression tree). The lack of consistent herbivore effects across sites most probably reflects differences in herbivore guilds, seasonality of herbivory, and variation in herbivore body-size distributions across sites, features for which data were not available. Larger, more detailed data sets will undoubtedly provide greater resolution of how different driver variables interact to influence mean woody cover.

These results have the power to inform savanna management strategies because they bear directly on our ability to predict savanna responses to changing environmental drivers. In particular, our data indicate that woody encroachment, a phenomenon in which many savannas across the world show a directional trend of increasing woody cover¹, may be a bounded process in savannas receiving a MAP of $<650 \pm 134$ mm, ultimately limited by water availability. For sites close to or at the MAP-controlled bound (Fig. 1), changes in precipitation regimes that lead to increased water availability⁶ therefore may be a cause for concern with respect to woody encroachment. However, the enormous variation in woody cover, with most sites far from the climatic bound (Fig. 1), suggests that processes other than MAP regulate actual tree cover in many savannas of Africa. In particular, our results suggest that if disturbances by fire, browsers

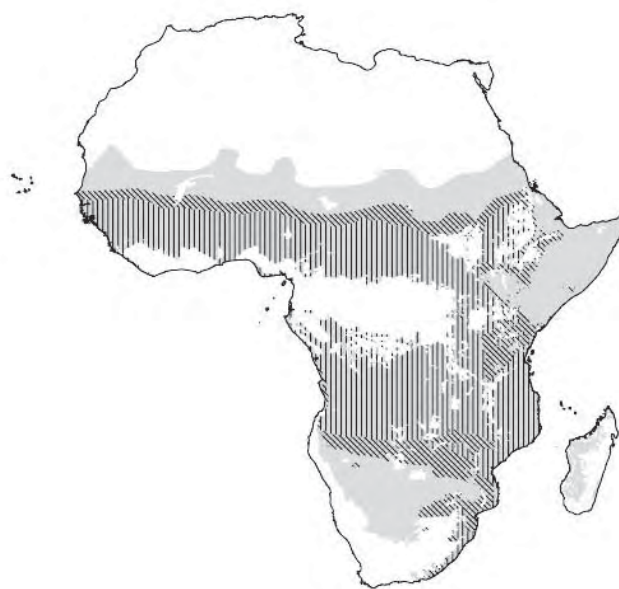


Figure 4 | The distributions of MAP-determined ('stable') and disturbance-determined ('unstable') savannas in Africa. Grey areas represent the existing distribution of savannas in Africa according to ref. 30. Vertically hatched areas show the unstable savannas (>784 mm MAP); cross-hatched areas show the transition between stable and unstable savannas (516–784 mm MAP); grey areas that are not hatched show the stable savannas (<516 mm MAP).

and humans were absent, then large sections of the African continent would switch to a wooded state (hatched regions in Fig. 4).

The patterns described here for African savannas suggest that the dominant ecological theories for tree–grass coexistence in these systems need to be combined: it is clear that most savannas are strongly affected by disturbances that maintain woody cover well below the resource-limited upper bound. Disturbance-based models do not consider and are unable to explain, however, the upper bound to tree cover. The results emerging from this continental scale analysis strongly indicate that water limits the maximum cover of woody species in many African savanna systems, but that disturbance dynamics control savanna structure below the maximum. These results have important implications both for our understanding of the fundamental nature of African savanna systems and for our ability to predict their responses to changing environmental drivers. It remains to be established whether the patterns observed here for African savannas also hold in other tropical savanna regions or in temperate savannas where the effects of winter precipitation and temperature on moisture distribution through the soil profile can markedly alter water partitioning between woody and herbaceous plants, and thus can influence maximum woody cover.

METHODS

Data collection. Data on projected woody cover (the percentage of ground surface covered when crowns are projected vertically), MAP, soil characteristics (texture, total nitrogen and phosphorus, and nitrogen mineralization), fire and herbivory regimes were gathered from several sources for a range of sites across Africa. We included only sites for which vegetation was sampled over sufficiently large spatial scales (>0.25 ha for plot measurements and >100 m for transect sampling). Sites located in riparian or seasonally flooded areas, or in net water run-on areas such as depressions, and sites in which trees were known to access ground water resources (that is, sources of water not dependent on rainfall in the immediate vicinity or in recent years) were excluded from the analysis because MAP is not a relevant descriptor of water availability in these sites. We also excluded sites that had been cultivated or harvested by humans <10 yr before sampling from the analysis.

Rainfall data included estimates from field measurements and regional rainfall maps ($n = 469$) and from fitted climatic grids (0.05° resolution,

$n = 383$) of monthly mean rainfall for Africa from the ANU-CRES (ref. 20; http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/santa_fe.html and <http://cres.anu.edu.au/outputs/africa.php>). Fire-return periods were obtained from field records ($n = 182$) and from burnt-area maps of Africa at 5-km resolution ($n = 670$) derived from AVHRR (advanced very high resolution radiometer) images based on 8 yr of data (1981–1983 and 1985–1991; ref. 21). Herbivore density estimates were available for 180 sites. Soils were obtained from 166 sites and analysed under standardized laboratory conditions for texture, total nitrogen and phosphorus, and nitrogen mineralization potential (see Supplementary Information). Our data set included sites encompassing a wide range of rainfall (132–1,185 mm MAP), fire-return intervals (1 to >50 yr), herbivore biomass (0–8,000 kg km⁻²), soil texture (sand, 6.7–98%; clay, 0.6–62.8%), soil percentage nitrogen (0.013–0.31%), soil total phosphorus (5–1,465 µg g⁻¹) and potential nitrogen mineralization rates (–22.8 to 153 µg g⁻¹ per week; see Supplementary Fig. S3).

Data analyses. To characterize the effects of MAP on the upper limit to woody cover across sites, we analysed data using a bent-cable form of a piece-wise linear model²² estimated with nonlinear quantile regression²³, as implemented in the 'quantreg' library in the statistical package R (<http://www.r-project.org/>). We used 0.90 to 0.99 conditional quantiles to obtain estimates near the upper boundary of the percentage of woody cover as it changes with MAP, which better reflects the process of MAP limiting maximum woody cover than does mean regression²⁴ (see Supplementary Information for details of this and additional analyses). We conducted additional analyses on the subset of sites that received <650 mm rainfall annually to investigate further how fire regimes, herbivory and soil properties influenced the upper bound on woody cover that was evident in these sites. We analysed these data by linear quantile regression²⁵, as implemented in the 'quantreg' library, which permits computation of confidence intervals for estimated parameters²⁶ and enabled us to test whether the regression slopes were different from zero.

In addition to analysing patterns in maximum woody cover, we used regression tree analysis²⁷, as implemented in the 'rpart' library in R, to determine how resource availability and disturbance regimes influenced mean realized woody cover in sites (see Supplementary Information). After tree construction, cross-validation procedures were used to prune trees to a size that best represented relationships that could be generalized outside the sample to the rest of the continent²⁸. Woody cover values were log-transformed to stabilize variances²⁸. To avoid problems arising from collinearity among soil variables, only sand content was retained for the analysis as it was the variable that was most strongly correlated to other soil variables (Supplementary Table S1). The results of the analysis were unchanged if grazer biomass and mixed feeder plus browser biomass were retained as two separate variables, or if total herbivore biomass was used as the predictor variable. The analysis was based on 161 sites for which data on MAP, fire-return intervals, herbivore biomass density and soil sand content were available. To determine whether the pruned tree explained more variance than a random tree of equal complexity, the square of the correlation coefficient (r^2) of the pruned tree was compared with r^2 values of similar sized trees generated from 2,000 random associations between predictor variables and woody cover²⁹. Further details on the methodology and results from additional analyses are provided in the Supplementary Information.

Received 26 April; accepted 22 July 2005.

- Scholes, R. J. & Archer, S. R. Tree–grass interactions in savannas. *Annu. Rev. Ecol. Syst.* **28**, 517–544 (1997).
- House, J. *et al.* Conundrums in mixed woody–herbaceous plant systems. *J. Biogeog.* **30**, 1763–1777 (2003).
- Jackson, R. B., Banner, J. L., Jobaggy, E. G., Pockman, W. T. & Wall, D. H. Ecosystem carbon loss with woody plant invasion of grasslands. *Nature* **418**, 623–626 (2002).
- Sankaran, M., Ratnam, J. & Hanan, N. P. Tree grass coexistence in savannas revisited—insights from an examination of assumptions and mechanisms invoked in existing models. *Ecol. Letters* **7**, 480–490 (2004).
- Jeltsch, F., Weber, G. E. & Grimm, V. Ecological buffering mechanisms in savannas: a unifying theory of long-term tree–grass coexistence. *Plant Ecol.* **150**, 161–171 (2000).
- Intergovernmental Panel on Climate Change. *Climate Change 2001: Synthesis Report. A Contribution of Working Groups I, II, and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds Watson, R. T., Core Writing Team) (Cambridge Univ. Press, Cambridge, UK, 2001).

- Simioni, G., Gignoux, J. & Le Roux, X. How does the spatial structure of the tree layer influence water balance and primary production in savannas? Results of a 3D modeling approach. *Ecology* **84**, 1879–1894 (2003).
- Sala, O. E. *et al.* Biodiversity—global biodiversity scenarios for the year 2100. *Science* **287**, 1770–1774 (2000).
- Bond, W. J., Midgley, G. F. & Woodward, W. I. The importance of low atmospheric CO₂ and fire in promoting the spread of grasslands and savannas. *Glob. Change Biol.* **9**, 973–982 (2003).
- Weltzin, J. F. & McPherson, G. R. (eds) *Changing Precipitation Regimes and Terrestrial Ecosystems: A North American Perspective* (Univ. Arizona Press, Tucson, 2003).
- Walter, H. *Ecology of Tropical and Subtropical Vegetation* (Oliver and Boyd, Edinburgh, 1971).
- Walker, B. H. & Noy-Meir, I. in *Ecology of Tropical Savannas* (eds Huntley, B. J. & Walker, B. H.) 556–590 (Springer, Berlin, 1982).
- Walker, B. H., Ludwig, D., Holling, C. S. & Peterman, R. M. Stability of semi-arid savanna grazing systems. *J. Ecol.* **69**, 473–498 (1981).
- Higgins, S. I., Bond, W. J. & Trollope, W. S. W. Fire, resprouting and variability: a recipe for grass–tree coexistence in savanna. *J. Ecol.* **88**, 213–229 (2000).
- Frost, P. G. *et al.* (eds) *Response of Savannas to Stress and Disturbance* (IUBS, Paris, 1986).
- van Langevelde, F. *et al.* Effects of fire and herbivory on the stability of savanna ecosystems. *Ecology* **84**, 337–350 (2003).
- Bond, W. J., Midgley, G. F. & Woodward, F. I. What controls South African vegetation—climate or fire? *S. Afr. J. Bot.* **69**, 79–91 (2003).
- McNaughton, S. J. The propagation of disturbance in savannas through food webs. *J. Veg. Sci.* **3**, 301–314 (1992).
- Walker, B. H. & Langridge, J. L. Predicting savanna vegetation structure on the basis of plant available moisture (PAM) and plant available nutrients (PAN): a case study from Australia. *J. Biogeog.* **24**, 813–825 (1997).
- Hutchinson, M. F., Nix, H. A., McMahon, J. P. & Ord, K. D. *Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling* (National Center for Geographic Information and Analysis, Santa Barbara, 1996).
- Barbosa, P. M., Stroppiana, D., Gregoire, J. M. & Pereira, J. M. C. An assessment of vegetation fire in Africa (1981–1991): burned areas, burned biomass, and atmospheric emissions. *Glob. Biogeochem. Cycles* **13**, 933–950 (1999).
- Chiu, G. S. *Bent-cable Regression for Assessing Abruptness of Change*. Thesis, Simon Fraser Univ. (2002).
- Koenker, R. & Park, B. J. An interior point algorithm for nonlinear quantile regression. *J. Econometrics* **71**, 265–283 (1994).
- Cade, B. S. & Noon, B. R. A gentle introduction to quantile regression for ecologists. *Front. Ecol. Environ.* **1**, 412–420 (2003).
- Koenker, R. W. & d'Orey, V. Computing regression quantiles. *Appl. Stat.* **36**, 383–393 (1987).
- Koenker, R. W. in *Asymptotic Statistics* (eds Mandl, P. & Huskova, M.) 349–359 (Springer, New York, 1994).
- Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. G. *Classification and Regression Trees* (Wadsworth International Group, Belmont, 1984).
- De'ath, G. & Fabricius, K. E. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**, 3178–3192 (2000).
- Rejwan, C., Collins, N. C., Brunner, L. J., Shuter, B. J. & Ridgway, M. S. Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology* **80**, 341–348 (1999).
- White, F. *The Vegetation of Africa: A Descriptive Memoir to Accompany the UNESCO/AETFAT/UNSO Vegetation Map of Africa* (UNESCO, Paris, 1983).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This paper arose from a workshop on savanna complexity funded by the NSF. We thank R. Boone, I. McHugh, R. Grant, H. Biggs, W. T. Starmer, P. M. Barbosa, D. Ruess, J. Rettenmayer, C. Williams, J. Klein, M. T. Anderson, W. J. Parton, J. C. Neff, N. Govender and the Kruger Park Scientific Services for comments, help with data collection and analysis, and for providing access to otherwise unpublished data.

Author Contributions All authors contributed data or intellectual input to the project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S. (mahesh@nrel.colostate.edu).

LETTERS

The phylogenetic position of the 'giant deer' *Megaloceros giganteus*

A. M. Lister¹, C. J. Edwards², D. A. W. Nock^{1†}, M. Bunce^{3†}, I. A. van Pijlen^{1,4}, D. G. Bradley², M. G. Thomas¹ & I. Barnes¹

The giant deer, or 'Irish elk', has featured extensively in debates on adaptation, sexual selection, and extinction. Its huge antlers—the largest of any deer species, living or extinct—formed a focus of much past work^{1–4}. Yet the phylogenetic position of the giant deer has remained an enigma. On the basis of its flattened antlers, the species was previously regarded as closely related to the living fallow deer^{5–7}. Recent morphological studies⁸, however, have challenged that view and placed the giant deer closer to the living red deer or wapiti. Here we present a new phylogenetic analysis encompassing morphological and DNA sequence evidence, and find that both sets of data independently support a sister-group relationship of giant and fallow deer. Our results include the successful extraction and sequencing of DNA from this extinct species, and highlight the value of a joint molecular and morphological approach.

With a fossil record extending from 400 kyr ago to its extinction about 8 kyr ago^{1–4}, the giant deer (*Megaloceros giganteus*) ranged from Ireland to central Siberia. Reaching a shoulder height of about 2 m and with antlers spanning up to 3.5 m, it was the largest known member of the 'Old World deer' (subfamily Cervinae)⁷.

The core taxa included in the study were the giant deer and its two putative alternative living sister-groups within the Cervinae, the fallow deer (European fallow, *Dama dama*, and Mesopotamian fallow, *D. mesopotamica*) and red deer (*Cervus elaphus*). To these were added the southeast Asian axis deer (*Axis axis*) and hog deer (*Axis porcinus*), in view of suggestions that *Axis* is a close living relative of fallow deer³. To broaden taxonomic sampling, wapiti (*C. canadensis*), sika (*C. nippon*) and Eld's deer (*C. eldi*) were added. The small muntjac deer of eastern Asia (*Muntiacus* spp.), which have been shown on morphological and molecular grounds to be basal living members of the Cervinae^{9–11}, were used as an

outgroup. In this way, every major clade of the Cervinae, identified in recent molecular studies¹¹, was represented.

A total of 988 base pairs (bp) of *M. giganteus* mitochondrial DNA (mtDNA) was obtained (see Methods) from two specimens of *M. giganteus* of widely divergent geographical origin. The first, from Ballynamintra Cave, Waterford, Ireland, has an uncalibrated AMS radiocarbon date of 11,567 ± 42 yr BP (KIA25446); the other, from Kamyshev Mire in western Siberia, has an uncalibrated date of 7,065 ± 38 yr BP (ref. 4). The sequences of the two specimens are 99.9% similar (one substitution) and so only the Siberian sequence was used in subsequent phylogenetic analyses. Homologous sequences were obtained for all the species in the study (Table 1).

Analysis of the mtDNA data under parsimony, likelihood, distance and bayesian criteria gave the same tree topology (Fig. 1a), and indicates a sister-group relationship between *M. giganteus* and *Dama*. To determine whether the molecular data could discriminate between the two previous morphologically derived hypotheses regarding the position of *Megaloceros*, a Kishino–Hasegawa (KH) test¹² was used to determine the significance of differences in likelihood values between on the one hand a topology in which *Megaloceros* forms a monophyletic group with species of *Dama*, and on the other a topology in which *Megaloceros* forms a monophyletic group with all or any species of the genus *Cervus*. Using the parameters derived above, a two-tailed KH test identifies the second of these as significantly less likely than the first, at $P < 0.01$.

A thorough examination of antlers, skulls, teeth and postcranial bones among the selected species led to a preliminary list of 250 variable characters. After discarding those with a high degree of intraspecific variability or difficulty of scoring (see Methods and Supplementary Information), the remaining set comprised 74 characters (Supplementary Table S1), of which 69 were phylogeneti-

Table 1 | Deer species and GenBank accession numbers of mtDNA sequences in this study

Taxon	Common name	ATP8	Control region	Cytochrome b
<i>Dama dama</i>	Fallow deer	AM072730	AF291895	AJ000022
<i>Dama mesopotamica</i>	Mesopotamian fallow deer	AM072731	AM072738	AM072742
<i>Axis axis</i>	Axis deer	AM072732	AM072739	AM072743
<i>Axis porcinus</i>	Hog deer	AM072733	AF291897	AY035874
<i>Cervus canadensis</i>	Wapiti	AM072737	AF058369	AF423199
<i>Cervus elaphus</i>	Red deer	AF104683	AF291886	AF423195
<i>Cervus nippon yezoensis</i>	Hokkaido Sika deer	AB108507	D50129	AB021095
<i>Cervus eldi</i>	Eld's deer	AM072734	AY137117	AY157735
<i>Muntiacus muntjak</i>	Indian muntjak	AY225986	AY225986	AY225986
<i>Muntiacus crinifrons</i>	Black muntjak	AY239042	AY239042	AY239042
<i>Megaloceros giganteus</i>	Giant deer (Russia)	AM072736	AM072740	AM072744
<i>Megaloceros giganteus</i>	Giant deer (Eire)	AM072736	AM072740	AM072745

¹Department of Biology, University College London, London WC1E 6BT, UK. ²Smurfit Institute of Genetics, Trinity College, Dublin 2, Ireland. ³Ancient Biomolecules Centre, Department of Zoology, Oxford University, Oxford OX1 3PS, UK. ⁴Department of Zoology, University of Leicester, Leicester LE1 7RH, UK. [†]Present addresses: MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, UK (D.A.W.N.); Department of Anthropology, McMaster University, Ontario L8S 4L9, Canada (M.B.).

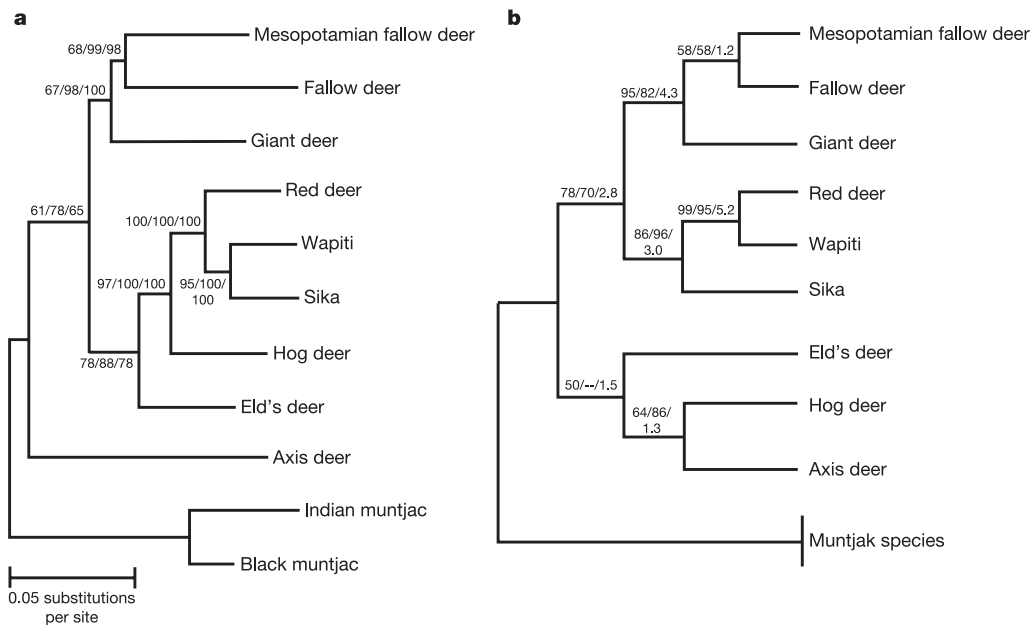


Figure 1 | Phylogenetic relationships among deer species based on molecular and morphological analyses. **a**, Maximum-likelihood phylogram constructed from 986–989 bp of mtDNA. Numbers above branches indicate, respectively, the percentage of trees that upheld that branch in an analysis of 1,000 bootstrap replicates, in a bayesian analysis of the molecular data set, and in a bayesian analysis of the combined

morphological and molecular data set. **b**, The single most parsimonious morphological cladogram. Tree length 123 steps, consistency index = 0.61, retention index = 0.58. Numbers above branches indicate, respectively, the percentage of trees (over 50%) in which a clade occurred in a maximum-parsimony analysis of 20,000 bootstrap replicates of the data, posterior probabilities from a bayesian analysis, and Bremer support indices.

cally informative. A single most-parsimonious tree was recovered, with strong statistical support for a monophyletic *M. giganteus*–*Dama* clade (Fig. 1b). Bayesian analysis produced a compatible but slightly less resolved tree, again with strong support for the *Megaloceros*–*Dama* grouping. The monophyletic clade of red deer, wapiti and sika was also recovered in both analyses, supporting the traditional view that red deer and wapiti are sister-species, in contrast to the wapiti-sika clade found in mtDNA (Fig. 1a; ref. 11). Axis, hog deer and Eld's deer were generally more basal on the tree but their topology was unstable. Neither our molecular analyses nor our morphological analyses provide evidence of a close relationship between *Axis* and the *Megaloceros*–*Dama* clade.

Combined analyses of the molecular and morphological data were undertaken with both bayesian and maximum-parsimony methods. The analyses consistently supported the *Megaloceros*–*Dama* and *Cervus elaphus*–*canadensis*–*nippon* clades. The bayesian topology was identical to the DNA-only tree (Fig. 1a), whereas using parsimony, detailed features such as the internal relationships of *Cervus*, depended on the relative weighting of morphological and molecular characters (Supplementary Information).

Eight derived morphological characters are unique to the *M. giganteus*–*Dama* clade in our data set (Supplementary Table S2); six of them occur in *M. giganteus* and both *Dama* species, and two in *M. giganteus* and *D. dama* only. They comprise two antler, one skull, two dental, one vertebral and two limb-bone characters (Fig. 2). Three further derived dental and one derived postcranial character are shared, apparently convergently, by the *Megaloceros*–*Dama* clade and one other species in the study. A final possible shared character, not included in the analysis, is the characteristic 'swollen larynx' ('Adam's apple') of *Dama*, seen in Palaeolithic representations of *M. giganteus*¹³ (Fig. 2a).

The sister-group relationship of *M. giganteus* and living *Dama*, shown by both our morphological and molecular data, corroborates the hypothesis of their relationship but must be viewed in the context of other Plio-Pleistocene fossil deer. *M. giganteus* is part of an array of well-documented species of 'giant deer' restricted to Eurasia, which

have been grouped as Tribe Megacerini Viret 1961 (refs 14–17). These share characters such as thickened mandible bones and, in most species, palmated antlers, although many authors^{15,18} have argued that they fall into two groups that might even have a biphyletic origin, one (*Praemegaceros*) derived from the Eurasian Plio-Pleistocene genus *Eucladoceros*, the other (including *M. giganteus*) of uncertain origin. However, a cladistic analysis¹⁹ found strong support for megacerine monophyly.

Our preliminary study of the other megacerine taxa indicates that the characters here identified as synapomorphies of *M. giganteus* and living *Dama* are common among them, but much less so among *Eucladoceros*, suggesting that *Dama* and all of the megacerine deer form a natural group. In this respect, we might invert the nineteenth-century appellation of *M. giganteus* as a 'giant fallow buck'²⁵ and consider the living fallow deer rather as the last representatives of a formerly speciose giant deer tribe. Although earlier forms have been claimed¹⁸, the first appearance of unquestionable megacerines is in the middle part of the Early Pleistocene epoch (about 1.4 Myr ago (ref. 16)), that of the *Dama dama*/*D. mesopotamica* lineage in the early Middle Pleistocene (about 700 kyr ago) and that of *M. giganteus*, plus related oriental forms, in the late Middle Pleistocene (about 400 kyr ago). This indicates the possibly rapid radiation of the clade in this relatively recent interval.

However, a series of European Pliocene to Early Pleistocene medium-sized cervids (spanning the approximate interval 3.0–1.0 Myr ago) have been regarded as forming a stem-group to modern *Dama* and have been placed in that genus by some authors^{8,20}. A recent cladistic analysis^{8,19} found derived characters linking these species and later *Dama*, although our preliminary observations on these taxa indicate that they lack most of the synapomorphies of modern *Dama* and *M. giganteus* identified here. This indicates homoplasy in the data. Either the entire clade of megacerines and modern *Dama* is of recent origin and postdates the split with the earlier 'Dama-like' forms, or else the latter are the sister-group of modern *Dama* and the split with megacerines is older. The latter would conform more closely to the deep divergence in mtDNA

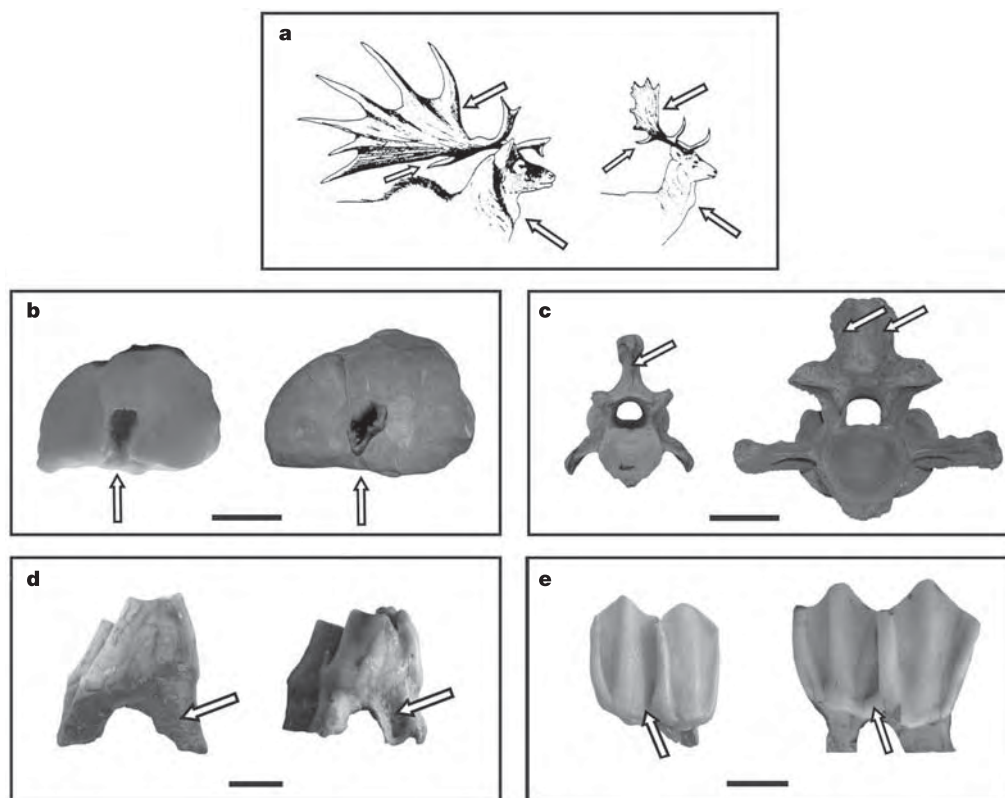


Figure 2 | Examples of morphological characters. **a**, *M. giganteus* (left) and *D. dama* (right), showing antler palmation, back tine, and expanded larynx. (Modified from Valerius Geist, *Deer of the World*, Stackpole Books, ref. 3.) **b**, Metacarpal of *C. canadensis* (left) and *M. giganteus* (right), showing separation of facets in *Cervus*. **c**, Posterior view of axis vertebra in *C. elaphus* (left) and *M. giganteus* (right), showing single medial ridge in *C. elaphus*, two

ridges bounding groove in *M. giganteus*. **d**, Posterior view of M^3 in *C. elaphus* (left) and *D. dama* (right), showing convex root in *C. elaphus*, concavity and labial ridge in *D. dama*. **e**, Labial view of upper molar of *C. elaphus* (left) and *M. giganteus* (right), showing horizontal basal ridges in *M. giganteus*. Scale bars, 2 cm (**b**), 5 cm (**c**) and 1 cm (**d**, **e**). See Supplementary Information for details of specimens.

sequence, with the *Dama*–*Megaloceros* split placed at 4–5 Myr ago (assuming a molecular clock and a divergence of the muntiacine and cervine deer at about 7 Myr ago¹¹). Resolving the relationships between these taxa, *Eucladoceros* spp., and other fossils sometimes implicated in megacerine ancestry^{17,18}, would enable us to explore palaeobiological questions such as the direction of size change, antler evolution, and associated adaptive issues. Some features of modern *Dama*, such as the robust axis vertebra (Fig. 2c) and relict parietal foramen²¹, suggest former adaptation to very large antler size and could correspond to ancestry from a large megacerine, but a comprehensive cladistic analysis of all these taxa is needed to address these issues further.

Last, the deep (7.85%) mtDNA divergence between European fallow deer and the endangered²² Mesopotamian fallow deer *Dama mesopotamica* highlights the distinctive nature of the latter taxon (corroborated by data in a recent study¹¹), in spite of its current demotion to a subspecies on the basis of similarities in gross morphology and in behaviour, and ease of cross-breeding in captivity^{9,23}. The morphological closeness of red deer and wapiti, to the exclusion of sika, is also notable, recalling earlier hypotheses of their relationship²⁴, but is in conflict with mtDNA data, inviting further investigation.

METHODS

Morphological analysis. Because of extensive intra-specific variability, all 250 original characters were scored on a series of individuals for each species. Variation was quantified as described previously²⁵, and characters were regarded as fixed only if they reached a high level of consistent expression (Supplementary Information). Intraspecific polymorphism was recoded as an ordered series with scaled weighting, and the data were analysed under parsimony by exhaustive

search with PAUP 4.0b10 (ref. 26), as well as by bayesian analysis (Supplementary Information).

Extraction and identification of ancient DNA. Sequence data were generated from about 0.1 g of cortical bone or tooth samples (Supplementary Information) with the use of established protocols for ancient DNA (University College London and the Ancient Biomolecules Centre, as in ref. 27; Trinity College Dublin, as in ref. 28).

In ancient mtDNA analysis, sequences can be recovered that are not authentic but derive from some external contaminant or the nuclear genome. We regard our *Megaloceros* sequences as genuine for the following reasons. First, entirely independently and without any exchange of materials, the two lead laboratories (University College London and Trinity College Dublin) generated near-identical (99.9%) sequences from different specimens obtained, directly in each case, from distant collections (in Eire and western Siberia respectively), by different workers (I.B. and C.J.E.) using different extraction methods and primer pairs. Second, by using separate samples of bone, a subset of sequences from the Kamyshev specimen (from primers amplifying ATP8, Cerv_cytb160F/288R, Cerv_cytb269F/403R and Cerv_cytb467F/604R) were independently replicated by M.B. at the Ancient Biomolecules Centre; and from the Ballynamindra specimen (using primers amplifying ATP8 and Cerv_cytb63F/176R, Cerv_cytb643F/760R and Cerv_cytb675F/786R) by I.B. at University College London. Third, the mtDNA sequences of the two *M. giganteus* specimens are clearly of cervid origin but are unique, sharing eight base substitutions not found in the other cervid taxa studied. Fourth, each sequence in the cytb contigs had a perfect match to overlapping fragments. Last, sequences of PCR product clones for several cytb fragments revealed a damage pattern characteristic for ancient DNA but not indicative of the presence of another contaminating sequence (Supplementary Information).

Molecular phylogenetic analyses. Between 986 and 989 characters (118 bp of ATPase 8, 113–116 bp of the control region, 755 bp of cytochrome *b*) were obtained for all taxa in the study (Table 1), of which 267 (27%) were variable and 158 (16%) were phylogenetically informative. We used PAUP 4.0b10 (ref. 26) for the initial analyses. Likelihood ratio tests supported the use of a GTR model (six

substitution types), incorporating site-specific rates for each of the four site categories (first, second, third and non-coding). For maximum-likelihood analyses, heuristic searches were conducted with starting trees generated by ten randomly derived stepwise addition sequences, with branch swapping by tree bisection-reconnection (TBR) and re-estimation of parameters. The maximum-likelihood topology (Fig. 1a) was also recovered by using both maximum parsimony and neighbour-joining with the HKY85 substitution model. Bootstrap support values were obtained by an analysis of 1,000 replicate data sets, with the use of maximum-likelihood analysis under a GTR + SS model with re-estimation of parameters at each step. Replicate data sets that maintained proportionality of each site type were generated with PAML v3.14 (ref. 29). Bayesian analyses of the morphological, molecular and combined data sets were conducted with MrBayes v3.1. Topology searches were initiated from random starting trees, with molecular data assigned a GTR + SS model, and morphological data analysed with distributed rates and likelihoods corrected for scoring bias caused by the presence of only variable characters in the data set. Both combined and molecular analyses were run for 5,000,000 generations, and the morphology-only analysis was run for 2,500,000 generations. Trees were sampled every 100 generations, with the first 25% discarded as burn-in³⁰.

Received 22 April; accepted 16 August 2005.

Published online 4 September 2005.

- Gould, S. J. The origin and function of 'bizarre' structures: antler size and skull size in the 'Irish Elk,' *Megaloceros giganteus*. *Evolution* **28**, 191–220 (1974).
- Moen, R. A., Pastor, J. & Cohen, Y. Antler growth and extinction of Irish Elk. *Evol. Ecol. Res.* **1**, 235–249 (1999).
- Geist, V. *Deer of the World: Their Evolution, Behavior and Ecology* (Swan Hill Press, Shrewsbury, 1999).
- Stuart, A. J., Kosintsev, P., Higham, T. F. & Lister, A. M. Pleistocene to Holocene extinction dynamics in giant deer and woolly mammoth. *Nature* **431**, 684–689 (2004).
- Lydekker, R. *The Deer of all Lands* (Rowland Ward, London, 1898).
- Millais, J. G. *The Mammals of Great Britain and Ireland* (Longman, Green & Co., London, 1906).
- Reynolds, S. H. *Monograph on the British Mammalia of the Pleistocene Period. The Giant Deer 1–58* (Palaeontographical Society Memoir, London, 1929).
- Pfeiffer, T. Die Stellung von *Dama* (Cervidae, Mammalia) im System plesiometa-carpaler Hirsche des Pleistozäns. *Cour. Forsch.-Inst. Senckenberg* **211**, 1–218 (1999).
- Groves, C. & Grubb, P. in *Biology and Management of the Cervidae* (ed. Wemmer, C. M.) 21–59 (Smithsonian Institution, Washington, D.C., 1987).
- Randi, E., Mucci, N., Pierpaoli, M. & Douzery, E. New phylogenetic perspectives on the Cervidae (Artiodactyla) are provided by the mitochondrial cytochrome *b* gene. *Proc. R. Soc. Lond. B* **265**, 793–801 (1998).
- Pitra, C., Fickel, J., Meijaard, E. & Groves, C. Evolution and phylogeny of Old World deer. *Mol. Phylogenet. Evol.* **33**, 880–895 (2004).
- Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170–179 (1989).
- Guthrie, R. D. *Frozen Fauna of the Mammoth Steppe* (Univ. of Chicago Press, Chicago, 1990).
- Lister, A. M. in *Biology and Management of the Cervidae* (ed. Wemmer, C. M.) 81–98 (Smithsonian Institution, Washington, DC, 1987).
- Azzaroli, A. & Mazza, P. On the possible origin of the Giant Deer genus *Megaceroides*. *Rend. Fis. Acc. Lincei* **3** (9), 23–32 (1992).
- Abbazzi, L. Remarks on the validity of the generic name *Praemegaceros* Portis 1920, and an overview on *Praemegaceros* species in Italy. *Rend. Fis. Acc. Lincei* **15** (9), 115–132 (2004).
- Croitor, R. & Kostopoulos, D. S. On the systematic position of the large-sized deer from Apollonia, Early Pleistocene, Greece. *Palaeontol. Z.* **78**, 137–159 (2004).
- Vislobokova, I. & Hu, C. On the evolution of Megacerines. *Vertebr. Palasiat.* **28**, 150–158 (1990).
- Pfeiffer, T. The first complete skeleton of *Megaloceros verticornis* (Dawkins, 1868) (Cervidae, Mammalia) from Bilshausen (Lower Saxony, Germany): description and phylogenetic implications. *Mitt. Mus. Naturk. Berl. Geowiss.* **5**, 289–308 (2002).
- Made, J. v. d. Ungulates from Atapuerca TD6. *J. Hum. Evol.* **37**, 389–413 (1999).
- Soergel, W. *Cervus megaceros mosbachensis* n. sp. und die Stammesgeschichte der Riesenhirsche. *Abh. Senck. Naturforsch. Ges.* **39**, 365–407 (1927).
- IUCN, *The IUCN Red List of Threatened Species* (IUCN, Gland, 2004).
- Vogler, U. New light upon the taxonomic status of European and Persian fallow deer on the basis of bone remains from Turkey. *Tierarztl. Praxis* **25**, 100–105 (1997).
- Lister, A. M. Evolutionary and ecological origins of British deer. *Proc. R. Soc. Edinb. B* **82**, 205–229 (1984).
- Lister, A. M. The morphological distinction between bones and teeth of fallow deer (*Dama dama*) and red deer (*Cervus elaphus*). *Int. J. Osteoarchaeol.* **6**, 119–143 (1996).
- Swofford, D. L. *PAUP* Phylogenetic Analysis using Parsimony (and Other Methods). Version 4.0b10* (Sinauer Associates, Sunderland, Massachusetts, 2002).
- Barnes, I., Matheus, P., Shapiro, B., Jensen, D. & Cooper, A. Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science* **295**, 2267–2270 (2002).
- MacHugh, D. E., Edwards, C. J., Bailey, J. F., Bancroft, D. R. & Bradley, D. G. The extraction and analysis of ancient DNA from bone and teeth: a survey of current methodologies. *Anc. Biomol.* **3**, 81–102 (2000).
- Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Appl. BioSci.* **13**, 555–556 (1997).
- Ronquist, F. & Huelsenbeck, J. P. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Currant, A. Friday, D. Hills, P. Jenkins, P. Kosintsev, L. Martin, N. Monaghan, T. Stuart, A. Vorobiev and E. Westwig for sampling and access to material; P. Grubb, D. MacHugh, C. O'hUigin and K. Wolfe for discussion; T. Burke and A. Cooper for laboratory facilities; P. Forey, J. Masters, A. Mitchell and M. Sánchez-Villagra for advice on cladistics; A. Murray and R. Rabinovich for technical assistance; and V. Geist and Stackpole Books for permission to reproduce the drawings in Fig. 2a. C.J.E. was supported by the Irish Research Council for Science, Engineering and Technology Basic Research Grant Scheme. I.A.vP was funded by BBSRC.

Author Information Sequences are deposited in GenBank under accession numbers AM072730–AM072749. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.B. (I.Barnes@ucl.ac.uk).

LETTERS

Morphine reward in dopamine-deficient mice

Thomas S. Hnasko¹, Bethany N. Sotak² & Richard D. Palmiter^{2,3}

Dopamine has been widely implicated as a mediator of many of the behavioural responses to drugs of abuse¹. To test the hypothesis that dopamine is an essential mediator of various opiate-induced responses, we administered morphine to mice unable to synthesize dopamine. We found that dopamine-deficient mice are unable to mount a normal locomotor response to morphine, but a small dopamine-independent increase in locomotion remains. Dopamine-deficient mice have a rightward shift in the dose–response curve to morphine on the tail-flick test (a pain sensitivity assay), suggesting either a decreased sensitivity to the analgesic effects of morphine and/or basal hyperalgesia. In contrast, dopamine-deficient mice display a robust conditioned place preference for morphine when given either caffeine or L-dihydroxyphenylalanine (a dopamine precursor that restores dopamine throughout the brain) during the testing phases. Together, these data demonstrate that dopamine is a crucial component of morphine-induced locomotion, dopamine may contribute to morphine analgesia, but that dopamine is not required for morphine-induced reward as measured by conditioned place preference.

Three decades of research have put the dopamine neurotransmitter system squarely at the centre of our understanding of the neural mechanisms through which drugs of abuse produce their effects^{1–3}. Indeed, it has been established that most drugs abused by humans increase midbrain dopamine neuron firing⁴ and/or dopamine release preferentially in the nucleus accumbens (NAc)⁵. Similarly, disruptions of the dopaminergic system, either pharmacologically or through brain lesions, can inhibit drug reward^{1,6–9}. Furthermore, animals will self-administer and acquire a conditioned place preference (CPP) for agents that increase dopamine receptor signalling (for example, agonists, dopamine transporter blockers) when delivered directly into the NAc¹. It has been proposed that the rewarding properties of opiates, such as heroin and morphine, are produced via activation of μ -opiate receptors located on GABAergic midbrain interneurons that negatively regulate dopamine cell firing¹⁰. Activation of these inhibitory G_{oi} -coupled μ -opiate receptors reduces the GABAergic tone onto midbrain dopamine neurons, thereby increasing their firing rate and the amount of dopamine released in the NAc. Indeed, animals will self-administer and display a place preference for opiates delivered directly into the ventral tegmental area^{11,12}. Moreover, mice lacking dopamine D₂ receptors fail to self-administer, or demonstrate a CPP for, morphine^{13,14} (but see also ref. 15). Although much evidence points to a key role for dopamine in mediating the effects of opiates, dopamine-independent mechanisms of opiate reward have been proposed⁹.

The dopamine-deficient (*Th*^{−/−}; *Dbh*^{Th/+}, see Methods) mice generated in our laboratory¹⁶ provide an ideal model to test whether dopamine is essential to the production of morphine-induced behavioural responses. These animals are severely hypoactive, hypophagic and require daily administration of L-dihydroxyphenylalanine (L-dopa) to prevent starvation. Therefore, to determine whether

dopamine is necessary for the acute effects of morphine, we administered morphine 18–24 h after L-dopa injection, when brain dopamine levels had fallen to <1% of control mice¹⁷.

Morphine, like most drugs of abuse, induces a locomotor response in rodents that is thought to reflect dopamine release in the striatum¹⁸. Administration of various doses of morphine to dopamine-deficient and control mice revealed that dopamine is a critical component of morphine-induced locomotion (Fig. 1a). At the highest dose, dopamine-deficient mice significantly increased locomotor activity but only to ~5% of control mice, indicating that morphine can stimulate a small amount of locomotion through a dopamine-independent pathway. In support of this conclusion, pre-treatment of dopamine-deficient mice with L-dopa (30 mg kg^{−1}) restored a robust locomotor response to morphine (Fig. 1b). Amphetamine pre-treatment, which purges any residual dopamine¹⁹, did not block the small morphine locomotor response in dopamine-deficient mice (Fig. 1c), providing further evidence that it is a dopamine-independent effect. Caffeine and other adenosine receptor antagonists stimulate locomotion when given to dopamine-deficient or control mice²⁰. The effects of caffeine pre-treatment on morphine-induced locomotion by dopamine-deficient mice were approximately additive (Supplementary Fig. 1a, b), indicating that generalized locomotor activation by caffeine does not replace the requirement for dopamine, and that a robust morphine locomotor response is dependent on, and specific to, dopamine release.

Because dopamine has been implicated in the modulation of pain sensitivity²¹ and certain types of morphine analgesia^{22,23}, we tested dopamine-deficient mice in the tail-flick task. Unlike many pain assays that require complex motor coordination, tail-flick is mediated by a spinal reflex arc, which dopamine-deficient mice are capable of performing. Dopamine-deficient mice displayed morphine analgesia, but the dose–response curve is shifted rightwards (Fig. 2a), suggesting that dopamine may mediate some of the analgesic effects of morphine. However, dopamine-deficient mice also displayed a faster tail-flick after saline injection, indicating that they may be more sensitive to pain under basal conditions. To confirm this hypothesis, we performed a tail-flick test at different temperatures. At every temperature tested, dopamine-deficient mice withdrew their tails more quickly (Fig. 2b). These data support the hypothesis that dopamine, perhaps originating from the A₁₁ dopamine neurons, which project directly to the spinal cord or through other indirect routes²¹, provides a tonic pain suppression signal that may be mediated through D₂ receptors²⁴.

Dopamine-deficient mice are bradykinetic and severely hypoactive and thus do not perform behavioural tasks that require voluntary movement. Therefore, we used viral gene transfer to restore dopamine synthesis to the caudate putamen. These virally rescued dopamine-deficient (vrDD) mice feed and locomote in the absence of L-dopa. We used vrDD mice in the CPP assay to test whether the hedonic experience of morphine was dependent on dopamine release outside of the caudate putamen (for example, the NAc). CPP is an

¹Graduate Program in Neurobiology & Behavior, ²Howard Hughes Medical Institute, and ³Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA.

assay whereby mice are repeatedly exposed to a treatment (for example, morphine) or vehicle in distinct chambers of a conditioning box and subsequently tested for their preference to occupy either chamber. During the drug-pairing sessions, the hedonic value

(reward) can become associated with the environment (dependent on associative learning), which is revealed during the testing phase. The testing phase also includes an incentive component—in the absence of motivation (drug seeking) the mouse may not manifest a preference. These *vrDD* mice readily explored the CPP box (Supplementary Fig. 2a, b) and displayed a strong preference for the morphine-paired chamber (Supplementary Fig. 2c). These results are consistent with previous lesioning experiments⁸ and suggest that dopamine is not required in brain regions outside the caudate putamen for the formation of morphine CPP.

To test whether morphine reward requires dopamine release anywhere in the brain, we developed a paradigm (Fig. 3a) that allows the dopamine-deficient mice to perform the exploratory aspects of the task yet receive the conditioning treatments during the dopamine-depleted state. Specifically, we gave a moderate dose of caffeine to dopamine-deficient or control mice during the baseline and testing phases of the CPP paradigm, but all pairing sessions (that is saline or morphine injections) were performed 18–24 h after L-dopa and in the absence of caffeine. This dose of caffeine induces locomotion and feeding in the absence of dopamine²⁰. Caffeine induced a sufficient locomotor response in dopamine-deficient mice that was ~45% of control (Fig. 3b, c). Despite the absence of dopamine during morphine administration, dopamine-deficient mice developed a CPP comparable to that of control mice at the two higher doses of morphine but not at the lowest dose (Fig. 3d). These data indicate that dopamine-deficient mice can develop a preference for the place where they receive morphine, demonstrating that dopamine release is not essential for this behaviour.

Because dopamine-deficient mice did not show a preference at the 2.5 mg kg⁻¹ dose of morphine, dopamine could be involved in some aspect(s) of morphine CPP. To ascertain whether dopamine is involved in the incentive (motivational or wanting) phase of

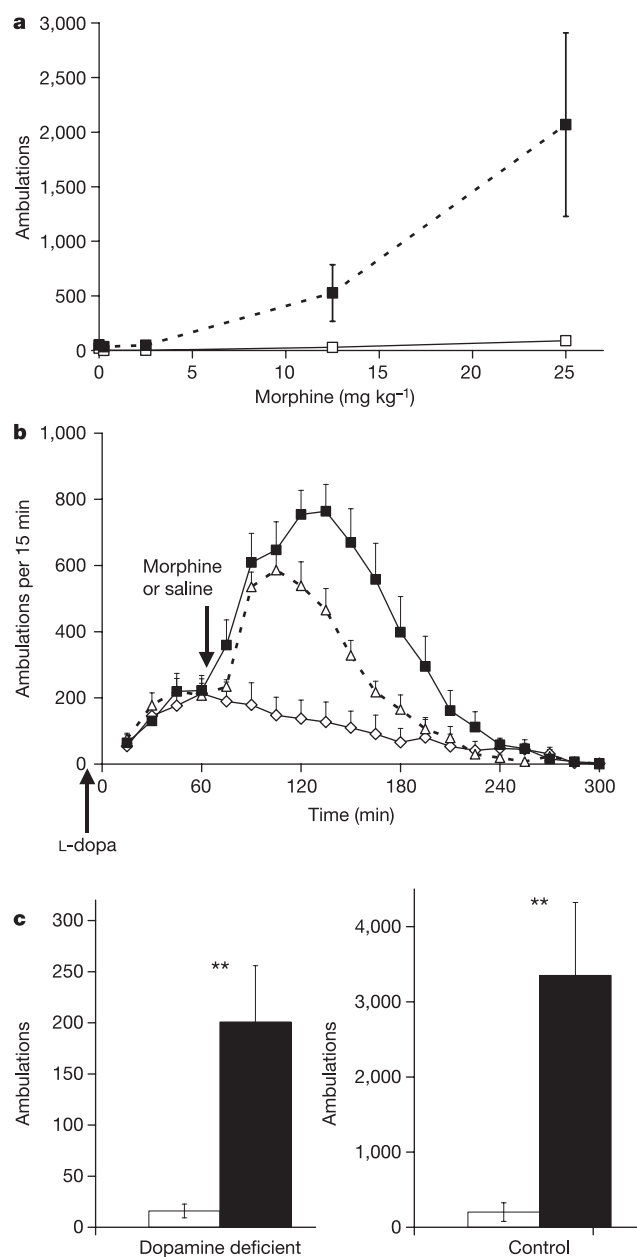


Figure 1 | Locomotor responses to morphine in dopamine-deficient and control mice. **a**, Cumulative 3-h morphine-induced locomotion by control ($n = 8$; filled squares) and dopamine-deficient mice ($n = 8$; open squares); dopamine-deficient mice have a severely blunted response. Repeated-measures analysis of variance (ANOVA); treatment effect $P < 0.001$; genotype effect $P < 0.05$; treatment–genotype interaction $P < 0.01$. **b**, Locomotor time course shows that 1 h L-dopa (30 mg kg⁻¹) pre-treatment of dopamine-deficient mice restores a robust locomotor response to morphine at 12.5 mg kg⁻¹ ($n = 8$; triangles) and 25 mg kg⁻¹ ($n = 8$; squares). Saline control shows the L-dopa response ($n = 8$; diamonds). Repeated-measures ANOVA on 3-h cumulative locomotion after morphine treatment; treatment effect $P < 0.001$. **c**, Amphetamine (3 mg kg⁻¹) pre-treatment does not eliminate subsequent morphine (25 mg kg⁻¹) response. Saline control, white bars; morphine, black bars. Repeated-measures ANOVA within genotype on 2-h cumulative locomotion after morphine treatment; treatment effect $**P < 0.01$. All data are presented as means $\pm$ s.e.m.

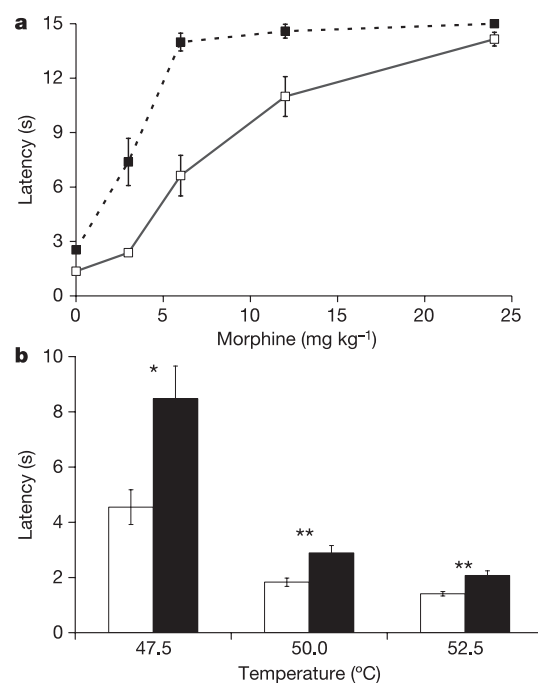


Figure 2 | Latencies to tail-flick by dopamine-deficient and control mice. **a**, Dose–response curve to morphine in control ($n = 8$; filled squares) and dopamine-deficient mice ($n = 9$; open squares). Repeated-measures ANOVA; treatment effect $P < 0.001$; genotype effect $P < 0.001$; treatment–genotype interaction $P < 0.001$. **b**, Heat response in dopamine-deficient ($n = 8$; white bars) or control mice ($n = 10$; black bars); dopamine-deficient mice are more sensitive to thermal stimulus. $*P < 0.05$, $**P < 0.01$; one-way ANOVA comparing genotypes within treatment. All data are presented as means $\pm$ s.e.m.

morphine CPP, we restored dopamine during the baseline and testing phases (by pre-treating the animals with a moderate dose of L-dopa instead of caffeine) but continued to conduct the pairing sessions in the absence of dopamine. This treatment rescued a CPP for morphine at 2.5 mg kg⁻¹ (Fig. 3e), suggesting that dopamine facilitates the manifestation of place preference (the incentive phase of CPP).

Our finding that dopamine-deficient mice are capable of learning to associate the hedonic effects of morphine with a particular environment is strong evidence that dopamine is not essential for mice to experience the rewarding effects of morphine or to learn to associate hedonic experiences with a specific environment. These conclusions are consistent with our previous work showing that

dopamine-deficient mice like sweets (sucrose or saccharin)²⁵ and can learn the location of food rewards without dopamine²⁶. This result suggests a number of possibilities regarding dopamine's role in mediating the acute rewarding properties of opiates. First, dopamine may not be involved in the acute reward produced by opiates. Indeed, several lines of evidence suggest that opiates can produce reward in the absence of increased dopamine release^{8,9}. Given the widespread expression of μ -opiate receptors, their activation in alternative brain regions such as the pedunculopontine tegmental nucleus²⁷ or NAC²⁸ may mediate the rewarding aspects of morphine independent of its effects on midbrain dopamine neuron firing. A second possibility is that dopamine contributes to morphine CPP but alternative neural pathways are of equal or greater importance. The fact that dopamine-deficient mice failed to develop a significant CPP at the lowest dose of morphine would be consistent with this possibility. However, because we were able to rescue CPP at the low dose of morphine by restoring dopamine only during the baseline and testing phases of the experiment, we conclude that dopamine facilitates the incentive (motivation) to seek the morphine-paired side of the conditioning chamber. A third possibility is that dopamine-independent compensatory mechanisms develop to mediate opiate reward in dopamine-deficient mice. We consider this unlikely because they are treated daily with L-dopa, which restores dopamine signalling for ~8 h each day; thus they are not chronically dopamine depleted. Furthermore, the dopamine-deficient mice are unable to compensate for other goal-directed behaviours (such as feeding) nor for the locomotor effects of morphine—although these behaviours can be rescued by acute restoration of dopamine via L-dopa.

The dopamine-deficient mice used in this study are uniquely suited to address questions about the importance of dopamine signalling throughout the brain for various biological processes, because we can maintain them with dopamine signalling (by providing L-dopa) and then study them in a dopamine-depleted state. This study demonstrates that although dopamine is a critical component of morphine-induced locomotion and modulates pain sensitivity, it is not required for mice to learn a conditioned association for morphine. We favour the hypothesis that dopamine is important for reward-seeking behaviour but is not essential for the hedonic experience of reward or for reward learning^{3,25,26}. We recognize, however, that on its own, this explanation fails to account for results such as the abolition of morphine CPP in rodents when co-administered with dopamine receptor antagonists⁷. However, these pharmacological studies are often difficult to interpret because many dopamine receptor antagonists can produce place-aversion⁷ and/or interfere with associative learning²⁹. In contrast, our explanation could account for the results of other experimental models (for example, lesions or genetic manipulations) in which dopamine signalling was disrupted and the animals failed to manifest reward—those manipulations may have blocked the ability of the animals to demonstrate a preference rather than experience the pleasure associated with morphine treatment.

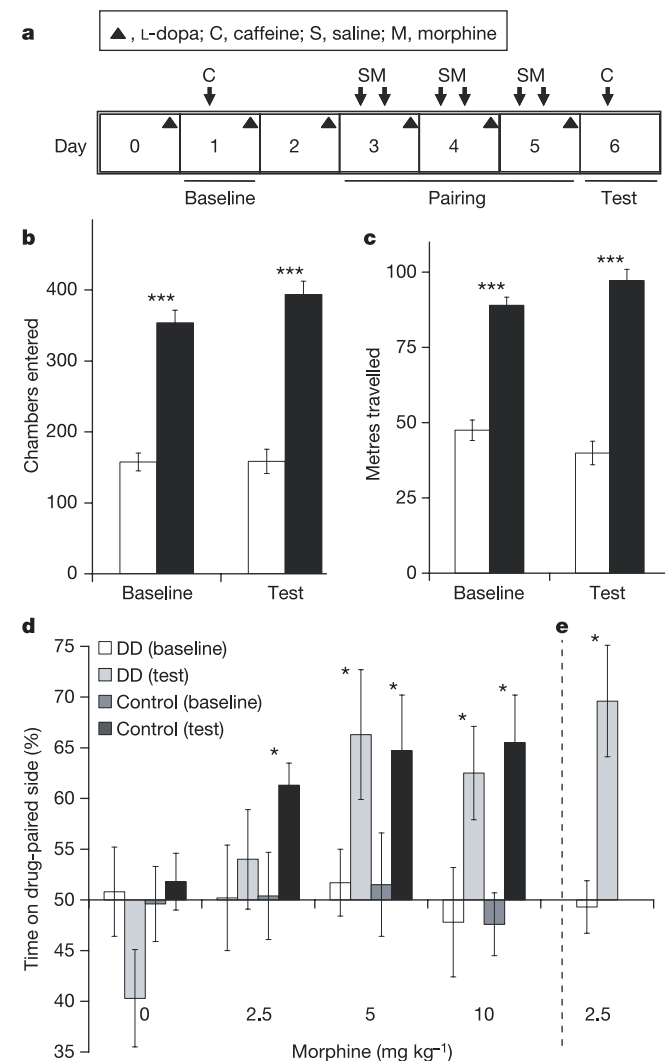


Figure 3 | Conditioned place preference in dopamine-deficient and control mice. **a**, Schematic illustrating the design of the CPP paradigm. **b**, **c**, Exploratory activity as measured by the number of chamber entries (**b**) or total distance travelled by dopamine-deficient ($n = 39$; white bars) or control mice ($n = 39$; black bars) during the baseline and testing phases of the caffeine CPP experiments (**c**). **d**, Dose-response to morphine (subcutaneously administered) in dopamine-deficient and control mice using caffeine as a locomotor stimulant ($n = 13, 10, 8$ and 8 for $0, 2.5, 5$ and 10 mg kg⁻¹ doses of morphine, respectively). DD, dopamine deficient. **e**, CPP for 2.5 mg kg⁻¹ morphine when using L-dopa (25 mg kg⁻¹) instead of caffeine during the baseline and testing phases ($n = 8$). Scores are presented as the percentage of time spent in the drug-paired side compared to the saline-paired side during the pre- and post-treatments; two-tailed paired t -test comparing time spent on drug-paired side before and after conditioning within genotypes and doses, * $P < 0.05$. All data are presented as means $\pm$ s.e.m.

METHODS

Subjects and treatments. Dopamine-deficient ($Th^{-/-}$; $Dbh^{Th/+}$) mice carrying two inactive tyrosine hydroxylase (Th) alleles, one intact dopamine β -hydroxylase allele (Dbh^{+}), and one Dbh allele with a targeted insertion of the Th gene (Dbh^{Th}) were created as described¹⁶. Controls included animals that carry at least one intact Th allele and one intact Dbh allele. Mice were maintained on a mixed C57BL/6 $\times$ 129/SvEv genetic background. All mice were housed under a 12/12-h light/dark cycle and temperature controlled environment with food (Purina, 5LJ5) and water available *ad libitum*. Except for L-dopa (see ref. 20), all drugs were dissolved in vehicle (PBS) and administered at a volume of 10 μ l g⁻¹. Mice were treated with PBS, morphine (RBI), amphetamine (Sigma), and/or caffeine (Sigma). All mice were treated in accordance with guidelines established by the National Institutes of Health and the University of Washington Animal Care Committee.

Behavioural assays. Locomotor studies were conducted in photo-beam activity cages as described²⁰. Dose-response was carried out in naive mice that received

escalating doses of morphine (0, 0.25, 2.5, 12.5, 25.0 mg kg⁻¹ body weight, intraperitoneally administered). All other locomotor assays were done using a Latin-square design such that each group of animals received each treatment. Tail-flick assays were performed using temperature-controlled water baths. Tails were dipped 0.5–1.0 cm beneath the water surface and the latency to withdrawal was measured with a 15-s cutoff. Each animal was tested in triplicate for each treatment and the average score was used. Escalating doses of morphine (0, 3, 6, 12, 24 mg kg⁻¹, intraperitoneally administered) were administered to naive mice 30 min before testing. A separate group of naive mice was used for the heat-response experiment. For CPP, we used two identical Plexiglas three-chamber boxes with two equal-sized compartments (20 × 20 cm) separated by a neutral grey chamber (20 × 7.5 cm). The boxes were separated by two sliding doors. The two large compartments had different coloured walls (black or white), different flooring (hard punched metal or stiff wire mesh), and were cued with different scents beneath the flooring (clove or ginger). These boxes were balanced such that, on average, mice tended to spend equal amounts of time in either chamber. All experiments were performed using the same schedule. On day 1 (18–24 h after L-dopa), mice were pre-treated with caffeine (15 mg kg⁻¹) 5 min before being placed in the centre chamber and allowed to explore the entire apparatus for 25 min. These sessions were recorded using a digital video camera (Sony). The next day, videos were analysed using Ethovision software (Noldus) to determine the time spent in each compartment, the total distance travelled, and the number of times an animal crossed from one chamber to another. Each animal was assigned one compartment for saline pairing and the other compartment for morphine pairing, such that there was minimal net difference in baseline times between compartments, within each genotype and dose (that is, unbiased paradigm). Any animal (and their littermate counterpart of the opposite genotype) that spent >65% of their time in a large compartment or >45% of their time in the neutral middle compartment were discarded from the study before assigning sides, due to the confounding factor of an endogenous preference or aversion. Conditioning was performed during days 3–5; each day each animal was treated with saline (subcutaneously administered) and confined to one side for 25 min in the morning, and treated with morphine (subcutaneously) and confined to the opposite side for 25 min in the afternoon. On day 6, animals were tested as on day 1. Data represent the per cent of time spent on the drug-paired side compared to the saline-paired side before and after conditioning (that is, the ratio does not include time spent in the centre chamber). For L-dopa rescue of CPP, the experiment was performed identically except that the mice were pre-treated with L-dopa (25 mg kg⁻¹) 30 min before baseline or testing instead of caffeine.

Received 1 July; accepted 1 September 2005.

- Spanagel, R. & Weiss, F. The dopamine hypothesis of reward: past and current status. *Trends Neurosci.* **22**, 521–527 (1999).
- Wise, R. A. Dopamine, learning and motivation. *Nature Rev. Neurosci.* **5**, 483–494 (2004).
- Berridge, K. C. & Robinson, T. E. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Brain Res. Rev.* **28**, 309–369 (1998).
- Gysling, K. & Wang, R. Y. Morphine-induced activation of A10 dopamine neurons in the rat. *Brain Res.* **277**, 119–127 (1983).
- Di Chiara, G. & Imperato, A. Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. *Proc. Natl Acad. Sci. USA* **85**, 5274–5278 (1988).
- Shippenberg, T. S., Bals-Kubik, R. & Herz, A. Examination of the neurochemical substrates mediating the motivational effects of opioids: role of the mesolimbic dopamine system and D-1 vs. D-2 dopamine receptors. *J. Pharmacol. Exp. Ther.* **265**, 53–59 (1993).
- Manzanedo, C., Aguilar, M. A., Rodriguez-Arias, M. & Minarro, J. Effects of dopamine antagonists with different receptor blockade profiles on morphine-induced place preference in male mice. *Behav. Brain Res.* **121**, 189–197 (2001).
- Pettit, H. O., Ettenberg, A., Bloom, F. E. & Koob, G. F. Destruction of dopamine in the nucleus accumbens selectively attenuates cocaine but not heroin self-administration in rats. *Psychopharmacology (Berl.)* **84**, 167–173 (1984).
- Bechara, A., Nader, K. & van der Kooy, D. A two-separate-motivational-systems hypothesis of opioid addiction. *Pharmacol. Biochem. Behav.* **59**, 1–17 (1998).
- Johnson, S. W. & North, R. A. Opioids excite dopamine neurons by hyperpolarization of local interneurons. *J. Neurosci.* **12**, 483–488 (1992).
- Bozarth, M. A. Neuroanatomical boundaries of the reward-relevant opiate-receptor field in the ventral tegmental area as mapped by the conditioned place preference method in rats. *Brain Res.* **414**, 77–84 (1987).
- Bozarth, M. A. & Wise, R. A. Intracranial self-administration of morphine into the ventral tegmental area in rats. *Life Sci.* **28**, 551–555 (1981).
- Maldonado, R. *et al.* Absence of opiate rewarding effects in mice lacking dopamine D2 receptors. *Nature* **388**, 586–589 (1997).
- Elmer, G. I. *et al.* Failure of intravenous morphine to serve as an effective instrumental reinforcer in dopamine D2 receptor knock-out mice. *J. Neurosci.* **22**, RC224 (2002).
- Dockstader, C. L., Rubinstein, M., Grandy, D. K., Low, M. J. & van der Kooy, D. The D2 receptor is critical in mediating opiate motivation only in opiate-dependent and withdrawn mice. *Eur. J. Neurosci.* **13**, 995–1001 (2001).
- Zhou, Q. Y. & Palmiter, R. D. Dopamine-deficient mice are severely hypoactive, adipsic, and aphagic. *Cell* **83**, 1197–1209 (1995).
- Szczypka, M. S. *et al.* Feeding behaviour in dopamine-deficient mice. *Proc. Natl Acad. Sci. USA* **96**, 12138–12143 (1999).
- Murphy, N. P., Lam, H. A. & Maidment, N. T. A comparison of morphine-induced locomotor activity and mesolimbic dopamine release in C57BL/6, 129/Sv and DBA/2 mice. *J. Neurochem.* **79**, 626–635 (2001).
- Heusner, C. L. *et al.* Viral restoration of dopamine to the nucleus accumbens is sufficient to induce a locomotor response to amphetamine. *Brain Res.* **980**, 266–274 (2003).
- Kim, D. S. & Palmiter, R. D. Adenosine receptor blockade reverses hypophagia and enhances locomotor activity of dopamine-deficient mice. *Proc. Natl Acad. Sci. USA* **100**, 1346–1351 (2003).
- Chudler, E. H. & Dong, W. K. The role of the basal ganglia in nociception and pain. *Pain* **60**, 3–38 (1995).
- Franklin, K. B. Analgesia and abuse potential: an accidental association or a common substrate? *Pharmacol. Biochem. Behav.* **59**, 993–1002 (1998).
- Altier, N. & Stewart, J. The role of dopamine in the nucleus accumbens in analgesia. *Life Sci.* **65**, 2269–2287 (1999).
- King, M. A., Bradshaw, S., Chang, A. H., Pintar, J. E. & Pasternak, G. W. Potentiation of opioid analgesia in dopamine₂ receptor knock-out mice: evidence for a tonically active anti-opioid system. *J. Neurosci.* **21**, 7788–7792 (2001).
- Cannon, C. M. & Palmiter, R. D. Reward without dopamine. *J. Neurosci.* **23**, 10827–10831 (2003).
- Robinson, S., Sandstrom, S. M., Denenberg, V. H. & Palmiter, R. D. Distinguishing whether dopamine regulates liking, wanting, and/or learning about rewards. *Behav. Neurosci.* **119**, 5–15 (2005).
- Bechara, A. & van der Kooy, D. The tegmental pedunculo-pontine nucleus: a brain-stem output of the limbic system critical for the conditioned place preferences produced by morphine and amphetamine. *J. Neurosci.* **9**, 3400–3409 (1989).
- Olds, M. E. Reinforcing effects of morphine in the nucleus accumbens. *Brain Res.* **237**, 429–440 (1982).
- Fenu, S., Bassareo, V. & Di Chiara, G. A role for dopamine D1 receptors of the nucleus accumbens shell in conditioned taste aversion learning. *J. Neurosci.* **21**, 6897–6904 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank our colleagues F. Perez for help generating mice, and E. Kremer for providing us with virus used for Supplementary Fig. 2. We thank the University of Washington NIDA center program for use of Noldus software. We thank our colleagues C. Chavkin, W. Watt and S. Luquet for comments on the manuscript. T.S.H. was supported in part by a grant from NIGMS.

Author Contributions The mouse model was developed in the R.D.P. laboratory. These experiments were designed and executed by T.S.H. with input from R.D.P. and assistance from B.N.S. for experiments shown in Fig. 1b and Supplementary Fig. 1.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.D.P. (palmiter@u.washington.edu).

LETTERS

BMP inhibition-driven regulation of *six-3* underlies induction of newt lens regeneration

Matthew W. Grogg^{1*}, Mindy K. Call^{1*}, Mitsumasa Okamoto², M. Natalia Vergara³, Katia Del Rio-Tsonis³ & Panagiotis A. Tsonis¹

Lens regeneration in adult newts is a classic example of how cells can faithfully regenerate a complete organ through the process of transdifferentiation^{1–6}. After lens removal, the pigment epithelial cells of the dorsal, but not the ventral, iris dedifferentiate and then differentiate to form a new lens. Understanding how this process is regulated might provide clues about why lens regeneration does not occur in higher vertebrates. The genes *six-3* and *pax-6* are known to induce ectopic lenses during embryogenesis^{7,8}. Here we tested these genes, as well as members of the bone morphogenetic protein (BMP) pathway that regulate establishment of the dorsal–ventral axis in embryos⁹, for their ability to induce lens regeneration. We show that the lens can be regenerated from the ventral iris when the BMP pathway is inhibited and when the iris is transfected with *six-3* and treated with retinoic acid. In intact irises, *six-3* is expressed at higher levels in the ventral than in the dorsal iris. During regeneration, however, only expression in the dorsal iris is significantly increased. Such an increase is seen in ventral irises only when they are induced to transdifferentiate by *six-3* and retinoic acid or by BMP inhibitors. These data suggest that lens regeneration can be achieved in noncompetent adult tissues and that this regeneration occurs through a gene regulatory mechanism that is more complex than the dorsal expression of lens regeneration-specific genes.

To determine the role of *six-3* and *pax-6* in the induction of transdifferentiation of the ventral iris, ventral iris cells were transfected in the presence or absence of retinoic acid with the appropriate constructs and examined for induction by using an *in vitro* transfection and *in vivo* transplantation system that reproduces the conditions seen *in vivo*^{10–12}. Retinoids have been shown to affect regeneration and to determine morphogenesis and differentiation of several tissues including the eye and limb^{13–16}. In addition, dorsal or ventral iris explants were treated with soluble BMP-4, BMP-7, chordin and a soluble competitor for the receptor BMPRI-A.

After transfection and implantation of aggregated pigment epithelial cells (PECs), scores of eyes were examined (see Supplementary Information). As a rule, untransfected dorsal PEC aggregates transdifferentiate to lens, whereas the ventral ones do not. Under the conditions outlined in the Methods, short-term culturing of cells does not interfere with the potential for lens transdifferentiation. Dorsal aggregates produced a lens in over 83% of tests (10/12), whereas the ventral ones, as expected, did not (0/11; Fig. 1a–c). It has been shown, through β -galactosidase staining, that the lens is indeed derived from the aggregate¹². Dorsal aggregates transfected with the constructs with retinoic acid treatment also transdifferentiated to lens (data not shown). With ventral PECs, however, only one protocol—namely, transfection of PECs with *six-3* in the presence

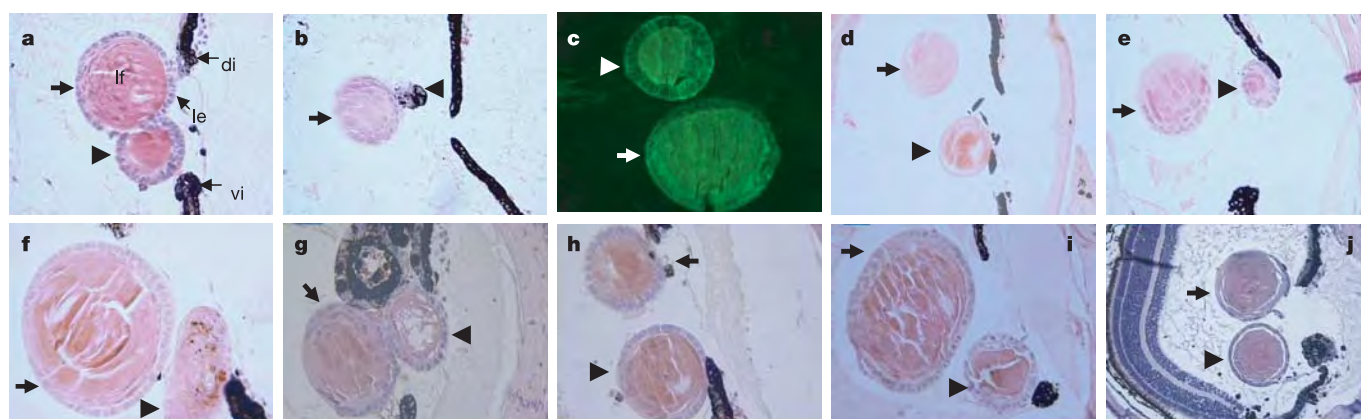


Figure 1 | Lens induction from ventral PECs. **a–f**, Lens induction by transplantation of PEC aggregates examined 30 d after transplantation. Thick arrows indicate the host regenerated lens; arrowheads indicate the PEC aggregate or the lens induced from the PEC aggregate. **a**, A control untransfected dorsal PEC aggregate that has transdifferentiated to lens. **le**, lens epithelium; **lf**, lens fibres; **di**, dorsal iris; **vi**, ventral iris. **b**, A control untransfected ventral PEC aggregate (arrowhead) that has remained

pigmented and failed to transdifferentiate to lens. **c**, Detection of crystallin synthesis in both a host lens and an induced lens with a lens fibre-specific antibody to β -crystallin²³. **d–f**, Induced lenses from ventral PEC aggregates transfected with *six-3* and treated with retinoic acid (RA). **g–j**, Induced lenses from ventral iris explants treated with BMPRI-A inhibitor (**g–i**) and with chordin (**j**).

¹Laboratory of Molecular Biology, Department of Biology, University of Dayton, Dayton, Ohio 45469-2320, USA. ²Division of Biological Science, Graduate School of Science, Nagoya University, Nagoya 464, Japan. ³Department of Zoology, Miami University, Oxford, Ohio 45056, USA.

*These authors contributed equally to this work.

of retinoic acid—led to the induction of lens transdifferentiation (Fig. 1d–f). This induction occurred at a rate (3/4; 75%) comparable to that seen in the dorsal aggregates. Neither treatment with retinoic acid alone nor transfection of ventral PEC cultures with *six-3* alone induced transdifferentiation.

In the BMP series we found that inhibition of the pathway by either the BMPR-IA competitor or chordin resulted in the induction of a lens from the ventral explants (3/15 and 1/8 respectively; Fig. 1g–j). The incidence of induction was low (17%); however, we regard this as highly significant because no untreated ventral explants differentiated to lens (0/27; 0% induction). This is in agreement with the established role of BMPs in maintaining ventral identity during embryogenesis and the fact that inhibiting the binding of BMPs to receptors results in dorsalization⁹. Notably, treatment of the dorsal iris explants with BMP-7, and to a lesser degree with BMP-4, considerably inhibited their ability to transdifferentiate to lens (1/12 (8.3%) and 5/12 (41.6%), respectively). Such results clearly indicate that BMPs maintain ventral identity and that inhibition of the pathway dorsalizes the ventral iris, allowing transdifferentiation.

To probe further the mechanism of induction, we undertook gene expression profiling of *six-3* and *BMPR-IA* during lens regeneration and during the experimental treatments that lead to the induction of lens regeneration from the ventral iris. Expression of *pax-6* was also assessed because of its known association with *six-3*. We examined samples of iris isolated 2, 4 and 8 d after lentectomy, as during this time dedifferentiation events that lead to regeneration from the dorsal iris have been initiated. Moreover, at later stages the vesicle starts expressing crystallins and differentiating to lens. Because these genes are also expressed in the differentiating lens, their induction-related expression might be ‘contaminated’. Several unexpected points emerged from the expression patterns that call for a revision of our view of the mechanism of lens regeneration.

First, both dorsal and ventral iris showed expression of all three genes. Analysis of the data to compare expression between the dorsal and ventral iris showed that the three genes were expressed more highly in the intact ventral iris. This pattern was maintained at day 8 but with a lesser relative fold change (Fig. 2a). However, analysis of the data to compare expression in the 2-, 4- and 8-d dorsal iris relative to the intact dorsal iris, and the 2-, 4- and 8-d ventral iris relative to the intact ventral iris, to correlate expression with the process of regeneration, revealed another pattern. The expression of *six-3* was increased only in the dorsal iris and seemed comparable at day 8 to that in the ventral iris. *BMPR-IA* and *pax-6* were also slightly upregulated (Fig. 2b–d). Upregulation of *six-3* in the dorsal iris started at day 4 (Fig. 2c), whereas that of *pax-6* and *BMPR-IA* started at day 8 (Fig. 2d). Thus, an increase in *six-3* seems to be important during the dedifferentiation process in the dorsal iris. Because regeneration occurs only from the dorsal iris and because the ventral iris also expresses these genes, our data suggest that gene regulation associated with the competency for lens regeneration aims to increase expression over a particular threshold and not simply to render a regulatory gene as dorsal-specific. Such a pattern for *six-3* is clearly shown when its expression at the different time points is directly compared with that in the intact dorsal iris (Fig. 2e).

Treatment of ventral iris cells with *six-3* and retinoic acid, which resulted in induction of transdifferentiation, showed a similar pattern of upregulation of *six-3*, *pax-6* and *BMPR-IA* when compared with the untransfected ventral cells (Fig. 2f). Treatment of the cells with retinoic acid alone or transfection of *six-3* alone, which did not induce the irises to differentiate to lens, did not show such a pattern (data not shown). Similarly, treatment of ventral iris explants with chordin, which also resulted in induction, invoked a marked upregulation of *six-3*, *pax-6* and, to a lesser degree, *BMPR-IA* in the treated ventral irises, as compared with the increase in untreated irises (Fig. 2g). *BMPR-IA* transcriptional regulation might not be that important for the induction. Notably, the rate of increase (as a relative fold change) in the treated ventral irises was comparable to

the increase in the regenerating 8-d dorsal iris. Thus, the treated ventral irises that transdifferentiated to lens adopted a gene expression profile (especially for *six-3*) that was seen only in the dorsal iris during dedifferentiation and regeneration. This observation indicates that when the ventral irises are coaxed to mimic patterns of regulatory events seen in the dorsal iris they become ‘dorsalized’ and thus can transdifferentiate into lens.

These expression patterns of *six-3* led us to examine whether there are subpopulations of cells in the dorsal or ventral iris that might account for these differences by using immunostaining to assess the distribution of Six-3-expressing cells. We stained serial sections along the nasal–temporal axis that spanned the whole iris (with distinct dorsal and ventral portions) and we counted the positive cells. Six-3-positive cells were found throughout the 8-d dorsal and ventral irises examined without apparent differences in their distribution (Fig. 3a), showing that the upregulation of *six-3* is not due to its expression in more cells. When the aggregates (or explants) transdifferentiated to lens, nearly all cells participated, also arguing against the existence of Six-3-expressing subpopulations (Fig. 1). We also examined the expression of Six-3 and BMPR-IA throughout the lens regeneration process. Figure 3b shows expression in dorsal and ventral iris at early stages (before vesicle formation) and Fig. 3c shows expression in

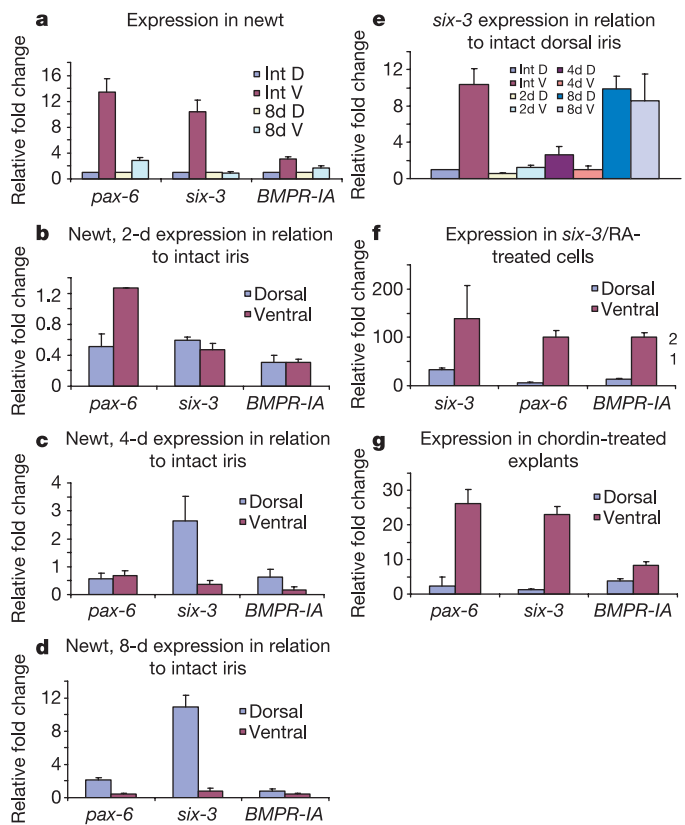


Figure 2 | Expression during lens regeneration and induction.

a, Comparison of *pax-6*, *six-3* and *BMPR-IA* expression between the newt dorsal and ventral iris. Expression is shown in intact irises and 8 d after lentectomy. Values in the dorsal iris have been set to 1 and those in the ventral iris are shown as relative fold changes. **b–d**, Comparison of *pax-6*, *six-3* and *BMPR-IA* expression in 2-, 4- and 8-d irises with that in intact irises. **e**, Comparison of *six-3* expression at all time points with that in the intact dorsal iris. **f**, Expression in PECs treated with *six-3* plus retinoic acid (RA) relative to untransfected cells. Note the considerable increase in exogenous *six-3* levels in the ventral PECs. Note that the levels of *pax-6* and *BMPR-IA* are much lower, as indicated by the numbers on the right y axis (used to accommodate the very high fold increase in *six-3*). **g**, Expression in chordin-treated iris explants relative to untreated explants. Data show mean $\pm$ s.d.

dorsal iris at later stages (in vesicles or regenerating lens). Expression in ventral iris at later stages was also positive (data not shown). The presence of *Six-3* in dorsal and ventral iris is consistent with the polymerase chain reaction (PCR) data; however, the immunostaining data are not quantitative.

We also considered whether the above 'treatments' are unique to the newt or whether they could induce lens transdifferentiation in irises from other vertebrates by testing another salamander, the axolotl, which possesses the ability to regenerate limbs and tail, but not the lens. None of the treatments induced transdifferentiation, indicating that the treatments are most probably newt-specific. However, it might be too premature to preclude the participation of *six-3* and BMP inhibition in lens regeneration in other species. The lens is regenerated from the dorsal iris in the newt but from the cornea in the pre-metamorphic frog—two strategies that differ from the embryonic induction of lens development and that could argue against absolute conservation of the inductive mechanisms.

To gain insight into the axolotl data, we examined gene expression in intact irises, in irises 8 d after lentectomy and in treated irises. In contrast to what was observed in the newt, the expression profiles in the axolotl differed considerably in both intact and 8-d irises (Fig. 4a). Expression in the intact and 8-d ventral irises was not higher than in the dorsal ones. Moreover, *six-3* was severely downregulated in the irises 8 d after lentectomy as compared with the intact irises. *pax-6* and *BMPR-IA* were slightly upregulated in both dorsal and ventral irises (Fig. 4b). This expression pattern is diametrically opposite to what we observed in the newt and indicates that negative regulation of *six-3*, as well as regulation in establishing thresholds, might be responsible for the lack of lens regeneration in the axolotl. So, why did the treatments fail to induce lens transdifferentiation? We think that it was because repression of *six-3* persisted even after the treatments. Indeed, treatment with chordin, which mediated upregulation of *six-3* and *pax-6* in the induced newt ventral irises, did not do so in the axolotl (compare Fig. 4c with Fig. 2g). Several explanations can account for the axolotl results. First, *six-3*

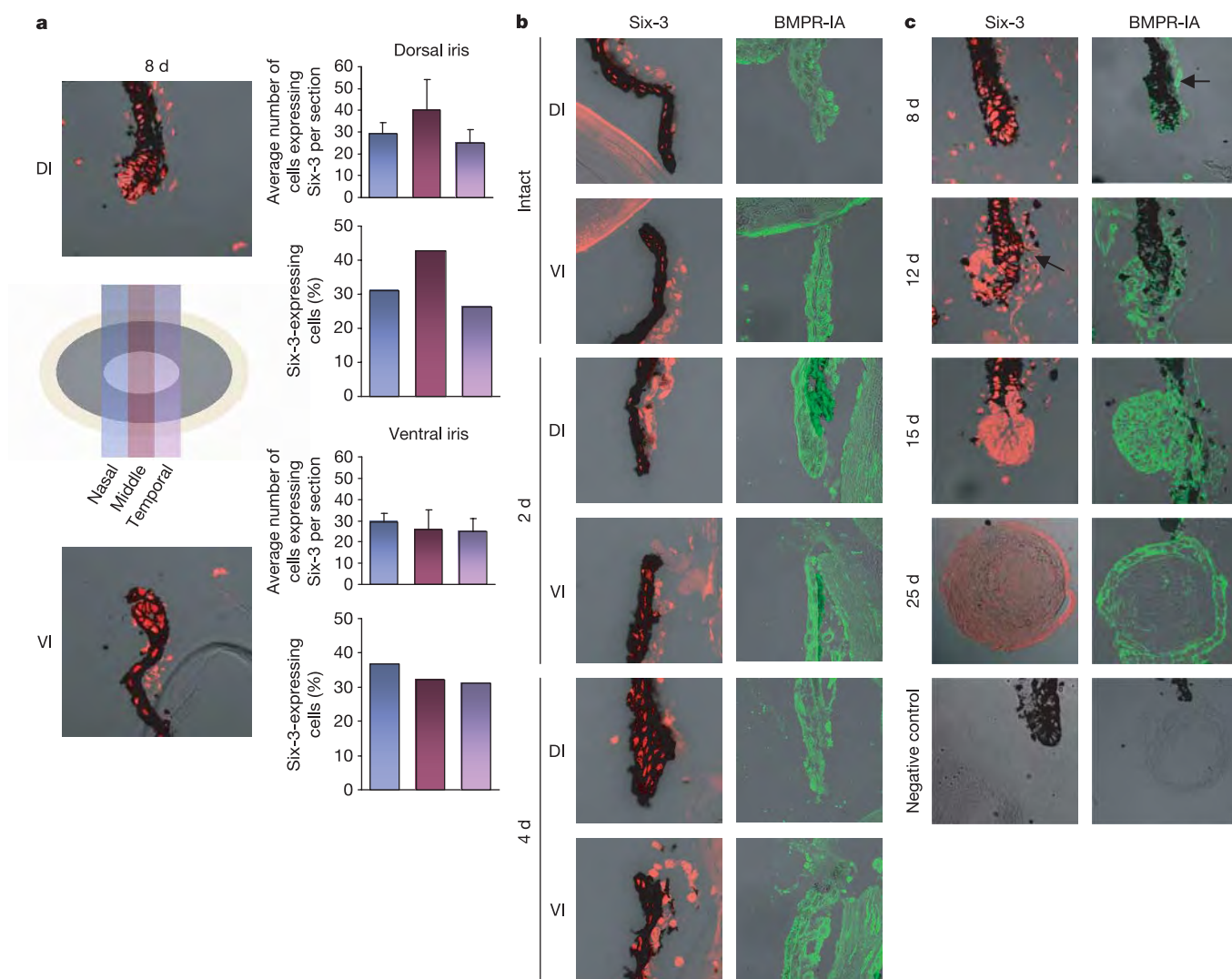


Figure 3 | Expression of *Six-3* and *BMPR-IA* during regeneration.

a, Identification of *Six-3*-positive cells in the dorsal (DI) and ventral (VI) iris 8 d after lentectomy, as assessed by immunostaining of serial sections along the nasal–temporal axis (see drawing). No apparent difference in cell numbers or distribution was seen (right). Graphical data are the mean $\pm$ s.d. **b**, Expression of *Six-3* and *BMPR-IA* in dorsal and ventral iris at early stages (before vesicle formation). Owing to heavy pigmentation at these early stages for the *BMPR-IA* staining, sections were bleached.

c, Expression of *Six-3* and *BMPR-IA* at later stages. At 12 d after lentectomy, a definite lens vesicle has been formed. The lens vesicle and the differentiating lens fibres are positive at 15 d, and by 25 d only cells in the lens epithelium express *Six-3*. The same patterns can be seen for *BMPR-IA*. Negative controls are shown at the bottom. All images shown are mergers of the dark field and DIC images. The stroma (arrows) always shows unspecific fluorescence.

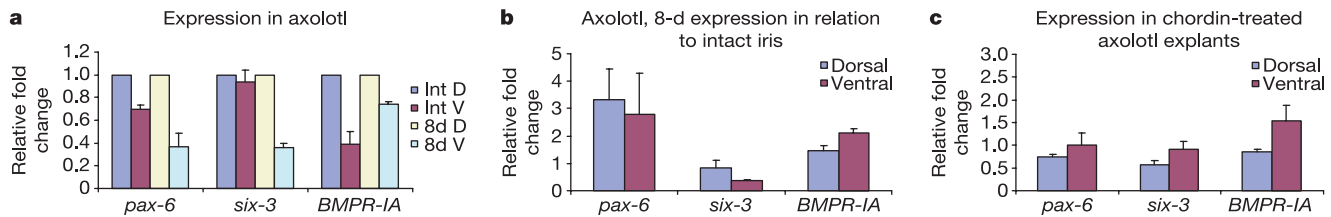


Figure 4 | Expression in axolotl. **a**, Comparison of *pax-6*, *six-3* and *BMPR-IA* expression between the axolotl dorsal and ventral iris. Expression is shown in intact irises and 8 d after lentectomy. Values in the dorsal iris have been set to 1 and those in the ventral iris are shown as relative fold

repression could be regulated more tightly in the axolotl than in the newt, possibly by an inhibitor of its transcription and function. Second, the axolotl PECs do not respond to these treatments equally and they may require optimized conditions. Third, the mechanism of induction of lens regeneration in different vertebrates follows unique pathways.

Notably, *pax-6* did not induce transdifferentiation of the ventral iris, even though it was shown to be upregulated in the chordin-treated ventral irises. However, *six-3* was not upregulated in *pax-6*-transfected cells (data not shown), which might be the reason why *pax-6* could not induce transdifferentiation. On the basis of other results from our laboratories, we now consider that *pax-6* is involved in later events of lens regeneration, such as the proliferation of PECs in both the dorsal and ventral iris and the control of crystallin synthesis (unpublished data). The fact that retinoic acid was also necessary for the induction most probably indicates that other factors regulated by retinoic acid are involved, a synergism that has been shown in other studies^{16–18}. Notably, both inhibitors of the BMP pathway and the *six-3*–*pax-6* loop are part of a network identified during induction of eye development¹⁹. The upregulation of *six-3* and *pax-6* in chordin-treated iris explants (Fig. 2g) suggests that the BMP signalling is upstream of the *six-3*–*pax-6* regulatory loop.

Previous work has shown that other important regulators of lens differentiation, such as FGFs, Sox2, MafB and members of the Hedgehog pathway, are expressed in both dorsal and ventral iris^{20,21}. This expression goes against the commonly held belief that regulatory genes involved in lens regeneration should be dorsal-specific. However, our detailed quantitative studies suggest that previously unknown regulatory events are involved in the induction of lens regeneration. Collectively, our data show that induction of lens regeneration can be achieved in noncompetent adult tissues—an important finding because ectopic lens formation has not been previously shown in adults and it opens new avenues in the field of vertebrate lens regeneration.

METHODS

All methods not listed here can be found in Supplementary Information.

Cloning of newt *six-3* and *BMPR-IA* partial cDNAs. *BMPR-IA* was cloned with RNA isolated from newt forelimb blastema (~2 weeks after amputation) using Tri Reagent (Molecular Research Center) according to the manufacturer's instructions. Dorsal PECs were used to clone a partial cDNA for *six-3*. We used 1 µg of RNA to synthesize cDNA using an iScript cDNA synthesis kit (BioRad). For PCR, a portion of the DNA was used along with *Taq* polymerase, 200 µM dNTPs and 800 nM primers. We used the following primers: *BMPR-IA*, forward (5'-TGCTGYATTGCTGAYTDDGG-3') and reverse (5'-GGRTCATYGGACCA-3'); *six-3*, forward (5'-CACTACCAGGAGGCCGAGAA-3') and reverse (5'-TCCTTGAAGCAGTGCGTCTT-3'). DNA was purified with a MinElute gel extraction kit (Qiagen). The fragment was cloned by the pGEM-T Easy vector system (Promega) and sequenced.

Immunostaining. Affinity-purified polyclonal antibodies were made against Six-3 and BMPR-IA peptides. Antibody against Six-3 was made in rabbit (New England Peptide) and antibody against BMPR-IA in chick (Cocalico). Newts were anaesthetized and the lens was removed through a slit in the cornea. The newts were killed at 2, 4, 8, 12, 15 and 25 d after lentectomy. The eyeballs were enucleated and fixed in 4% formaldehyde for 4 h, washed in 1 × PBS buffer,

changes. **b**, Comparison of expression in 8-d irises with intact irises.

c, Expression in chordin-treated iris explants relative to untreated explants. Data show mean ± s.d.

cryoprotected in 30% sucrose, embedded in OCT (Andwin Scientific), frozen and sectioned at 10 µm. Slides with frozen serial sections were washed several times in PBS and 1% saponin (Sigma) and incubated in 10% goat serum in PBS. Occasionally, to reduce pigmentation, sections were bleached in 0.1% potassium permanganate for 10 min, immersed in 0.5% oxalic acid for 5 min and then rinsed in PBS.

The samples were incubated at 4 °C overnight with primary antibody against newt Six-3 diluted 1:10 in blocking solution or against newt BMPR-IA diluted 1:100 in blocking solution, washed in 0.3% PBST and PBS, and then incubated with the following secondary antibody for 2 h at 37 °C: Alexafluor 546-conjugated goat anti-rabbit (Molecular Probes) for Six-3, or FITC-conjugated rabbit antibody against chicken (Sigma) for BMPR-IA, diluted 1:200 in 10% goat serum in PBST. The sections were washed with PBST and PBS, and covered with coverslips using Vectashield (Vector Labs). Images were taken by confocal microscopy.

Real-time PCR. RNA was isolated from iris tissue and PECs by using Tri Reagent (Molecular Research Center) according to the manufacturer's instructions. We used the following tissues and cells: intact dorsal and ventral iris; 2-, 4-, 8-d dorsal and ventral iris; cells isolated from dorsal and ventral iris; dorsal and ventral iris cells transfected with *six-3* and treated with retinoic acid, transfected with *six-3* alone, or treated with retinoic acid alone; explants from dorsal and ventral iris; dorsal and ventral iris explants treated with chordin or BMPR-IA; axolotl intact and 8-d dorsal and ventral iris; and chordin-treated axolotl dorsal and ventral iris. The RNA isolated was used to evaluate expression of *six-3*, *BMPR-IA* and *pax-6* (along with a suitable reference gene) by real-time PCR, as well as by RT-PCR to verify that the correct fragment was amplified. Appropriate negative controls were included in all sets.

We used 0.75 µg of RNA to synthesize cDNA using an iScript cDNA synthesis kit (BioRad). All real-time PCRs were done with an iCycler (BioRad). For each real-time PCR reaction run in triplicate, 2 µl of cDNA, 800 nM primers and iQ SYBR Green Supermix (BioRad) were used. Primers were designed from the cloned cDNAs for *six-3* and *BMPR-IA* and from a published sequence of *pax-6*. The new primers were *rpL27*, forward (5'-TACAACCACTTGATGCCA-3') and reverse (5'-CAGTCTTGATCGTTCCTCA-3'); *pax-6*, forward (5'-CTGGGCAGGTATACGAG-3') and reverse (5'-GTCTCTGATTCCAGGC-3'); *six-3*, forward (5'-CAAGAAGTTCCTGCTGC-3') and reverse (5'-GGTAGGGTCTCTGATGAC-3'); *BMPR-IA*, forward (5'-TGCTGTATTGCTGATTAGG-3') and reverse (5'-ATAGGTATCAAAGCAGTCCA-3'). The axolotl primers were *RP*, forward (5'-CATCAGATCAAGCAAGCAGTA-3') and reverse (5'-CCAATGCAGCAGTTAGATG-3'); *pax-6*, forward (5'-GAGTGCTCCGC AACCTG-3') and reverse (5'-ATTCGTGTTCTCGCCTCC-3'); *six-3* (same as newt); *BMPR-IA*, forward (5'-CAGTGCTGATTGCTGAT-3') and reverse (5'-GGCTACTTCCCAATAACC-3').

For each real-time PCR the basic program was as follows: denaturation at 95 °C, annealing at 50.6 °C and extension at 72 °C (40 cycles). To minimize the background caused by primer-dimer formation, an extra step was added (78 °C for 6 s) at the end of each cycle. The readings were taken during this step. Data analysis was done with the Pfaffl method²². The reference genes were *rpL27* for newt and *RP* for axolotl, which both encode ribosomal proteins.

Received 25 March; accepted 25 August 2005.

- Colucci, V. L. Sulla rigenerazione parziale dell'occhio nei Tritoni-Istogenesi e sviluppo. Studio sperimentale. *Mem. R. Acad. Sci. 1st Bologna Ser.* 51, 593–629 (1891).
- Wolff, G. Entwicklungsphysiologische Studien. I. Die regeneration der urodelenlinse. *Wilhelm Roux Arch. Entwickl.-Mech. Org.* 1, 380–390 (1895).
- Tsonis, P. A. Regeneration in vertebrates. *Dev. Biol.* 221, 273–284 (2000).
- Eguchi, G. Electron microscopic studies on lens regeneration I. Mechanism of depigmentation of the iris. *Embryologia* 8, 45–62 (1963).

5. Eguchi, G. Electron microscopic studies on lens regeneration. II. Formation and growth of lens vesicle and differentiation of lens fibers. *Embryologia* **8**, 247–287 (1964).
6. Yamada, T. Control mechanisms in cell-type conversion in newt lens regeneration. *Monogr. Dev. Biol.* **13**, 1–126 (1977).
7. Oliver, G., Loosli, F., Koster, R., Wittbrodt, J. & Gruss, P. Ectopic lens induction in fish in response to the murine homeobox gene *six-3*. *Mech. Dev.* **60**, 233–239 (1996).
8. Altmann, C. R., Chow, R. L., Lang, R. A. & Hemmati-Brivanlou, A. Lens induction by Pax-6 in *Xenopus laevis*. *Dev. Biol.* **185**, 119–123 (1997).
9. DeRobertis, E. M. & Kuroda, H. Dorsal–ventral patterning and neural induction in *Xenopus* embryos. *Annu. Rev. Cell Dev. Biol.* **20**, 285–308 (2004).
10. Okamoto, M., Ito, M. & Owaribe, K. Difference between dorsal and ventral iris in lens producing potency in normal lens regeneration is maintained after dissociation and reaggregation of cells from the adult newt, *Cynops pyrrhogaster*. *Develop. Growth Differ.* **40**, 11–18 (1998).
11. Ito, M., Hayashi, T., Kuroiwa, A. & Okamoto, M. Lens formation by pigmented epithelial cell reaggregate from dorsal iris implanted into limb blastema in the adult newt. *Dev. Growth Differ.* **41**, 429–440 (1999).
12. Hayashi, T. et al. Highly efficient transfection system for functional gene analysis in adult amphibian lens regeneration. *Develop. Growth Differ.* **43**, 361–370 (2001).
13. McCaffery, P., Lee, M., Wagner, M. A., Sladek, N. E. & Drager, U. C. Asymmetrical retinoic acid synthesis in the dorsoventral axis of the retina. *Development* **115**, 371–382 (1992).
14. Tsonis, P. A. *Limb Regeneration* (Cambridge Univ. Press, Cambridge, 1996).
15. Tsonis, P. A., Trombley, M. T., Rowland, T., Chandraratna, R. A. S. & Del Rio-Tsonis, K. Role of retinoic acid in lens regeneration. *Dev. Dyn.* **219**, 588–593 (2000).
16. Maden, M. & Hind, M. Retinoic acid, a regeneration-inducing molecule. *Dev. Dyn.* **226**, 237–244 (2003).
17. Lee, S. H., Fu, K. K., Hui, J. N. & Richman, J. M. Noggin and retinoic acid transform the identity of avian facial prominences. *Nature* **414**, 909–912 (2001).
18. Skillington, J., Choy, L. & Derynck, R. Bone morphogenetic protein and retinoic acid signalling cooperate to induce osteoblast differentiation of pre-adipocytes. *J. Cell Biol.* **159**, 135–146 (2002).
19. Zuber, M. E. et al. Specification of the vertebrate eye by a network of eye field transcription factors. *Development* **130**, 5155–5167 (2003).
20. Hayashi, T. et al. FGF2 triggers iris-derived lens regeneration in newt eye. *Mech. Dev.* **121**, 519–526 (2004).
21. Tsonis, P. A. et al. A novel role of the hedgehog pathway in lens regeneration. *Dev. Biol.* **267**, 450–451 (2004).
22. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res.* **29**, e45 (2001).
23. Sawada, K., Agata, K., Yoshiki, A. & Eguchi, G. A set of anti-crystallin monoclonal antibodies for detecting lens specificities: β -crystallin as a specific marker for detecting lentoidogenesis in culture of chicken lens epithelial cells. *Japn. J. Ophthalmol.* **37**, 355–368 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Hayashi for sharing his expertise with the culturing methods for PECs; G. Oliver and D. Englecamp for the *six-3* and *pax-6* expression vectors, respectively; members of our laboratories, M. Metheney, A. Thitoff, S. Gainer, M. Tsavaris, M. Tubb and M. Sander for help with histology; V. Mitashov and E. Makarev for providing information on *P.Waltl six-3* sequences; and E. Tanaka for information on axolotl *BMPR-IA* sequences. This work was supported by a grant from the NIH (to P.A.T.).

Author Information Sequences for *six-3* and *BMPR-IA* have been deposited in GenBank under accession numbers AY799802 and AY795966, respectively. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.A.T. (panagiotis.tsonis@notes.udayton.edu).

A colonization factor links *Vibrio cholerae* environmental survival and human infection

Thomas J. Kirn^{1*†}, Brooke A. Jude^{1*} & Ronald K. Taylor¹

Many bacteria that cause diseases must be able to survive inside and outside the host. Attachment to and colonization of abiotic or biotic surfaces is a common mechanism by which various microorganisms enhance their ability to survive in diverse environments¹. *Vibrio cholerae* is a Gram-negative aquatic bacillus that is often found in the environment attached to the chitinous exoskeletons of zooplankton^{2,3}. It has been suggested that attachment to zooplankton enhances environmental survival of *Vibrio* spp., probably by providing both an abundant source of carbon and nitrogen and protection from numerous environmental challenges⁴. On ingestion by humans, some serogroups of *V. cholerae* cause the diarrhoeal disease cholera⁵. The pathophysiology of cholera is a result of the effects of cholera toxin on intestinal epithelial cells. For sufficient quantities of cholera toxin to reach the intestinal epithelium and to produce clinical symptoms, colonization of the small bowel must occur. Because most *V. cholerae* do not colonize humans, but all probably require strategies for survival in the environment, we considered that colonization factors selected for in the environment may be the same as those required for intestinal colonization of humans. In support of this hypothesis, here we have identified a single protein required for efficient intestinal colonization that mediates attachment to both zooplankton and human epithelial cells by binding to a sugar present on both surfaces.

To identify proteins involved in *V. cholerae* intestinal colonization, we designed an *in vitro* cellular attachment screen to isolate *V. cholerae* mutants deficient for epithelial cell attachment (see the Methods and Supplementary Fig. S1). We identified several mutants that showed a reduced attachment phenotype. One of these mutants showed this defective phenotype owing to a loss of normal flagellar function, whereas several others were found to be sensitive to Triton, resulting in an apparent, but false, attachment defect in our assay. From among the mutants, one transposon insertion resulting in an attachment defect was determined to lie in an open reading frame that is predicted to encode a putative chitinase (VCA0811)⁶.

V. cholerae can grow in the presence of chitin as a sole source of carbon, and several *V. cholerae* proteins with chitinase and chitin-binding activity have been described^{7–10}. A previously isolated chitin-binding activity has an apparent molecular mass of 53 kDa and has been shown to be inhibited by the chitin subunit *N*-acetylglucosamine (GlcNAc)^{9–11}. Because VCA0811 encodes a putative 53-kDa protein, we thought that the gene product of VCA0811 might be identical to the previously described GlcNAc-sensitive chitin-binding protein. We therefore named the VCA0811 gene product GlcNAc-binding protein A and the gene *gbpA*.

A deletion of *gbpA* concomitant with fusion of *lacZ* to the native *gbpA* promoter to provide a transcriptional readout was constructed on the wild-type *V. cholerae* chromosome (Δ *gbpA*). The resulting

mutant was defective in attachment to the same extent as the original transposon insertion, was not deficient for *in vitro* growth as compared with wild type, and produced wild-type levels of TcpA (Supplementary Figs S2 and S3), the pilin subunit of the potent *V. cholerae* colonization factor toxin coregulated pilus (TCP). The Δ *gbpA* mutant was complemented with a plasmid bearing a *GbpA*–His fusion (Δ *gbpA*, p*GbpA*–His). We confirmed the absence of *GbpA* production in the Δ *gbpA* mutant and the restoration of *GbpA* production in the Δ *gbpA*, p*GbpA*–His strain by immunoblotting (Supplementary Fig. S4). These strains were then tested in the epithelial cell attachment assay. Δ *gbpA* attachment was reduced by 50% as compared with wild type, representing a significant difference in attachment between wild type and Δ *gbpA* (Fig. 1). Overexpression of p*GbpA*–His in the mutant background restored wild-type levels of attachment.

On the basis of the significant chitin-binding homology present in *GbpA*, we considered that the mechanism by which *GbpA* mediates epithelial cell attachment was related to the protein's chitin-binding activity. Although chitin is not present on the epithelial surface, the chitin monomer, GlcNAc, is a common modification of glycoproteins and lipids present on the intestinal epithelium^{12,13}. To determine whether *GbpA* can mediate *V. cholerae* attachment to chitin and GlcNAc, wild-type, Δ *gbpA* and Δ *gbpA*, p*GbpA*–His *V. cholerae*, all bearing a plasmid expressing the green fluorescent protein (pGreen–Tir¹⁴), were incubated with chitin beads or GlcNAc-coated beads. Wild-type *V. cholerae* bound to the surface of the chitin beads and GlcNAc beads in greater numbers than did the Δ *gbpA* mutant (Fig. 2a, b). The Δ *gbpA*, p*GbpA*–His strain was intermediate between wild type and the Δ *gbpA* mutant in chitin and GlcNAc bead attachment. These results confirm that *GbpA* has a role in *V. cholerae*

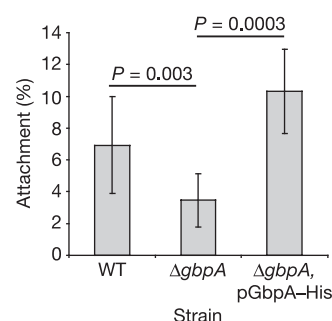


Figure 1 | VCA0811 encodes a putative chitin-binding protein (*GbpA*) required for epithelial cell attachment. Shown is the number (mean $\pm$ s.d.) of wild-type, Δ *gbpA* and Δ *gbpA*, p*GbpA*–His *V. cholerae* attached to HT-29 epithelial cells (divided by input number; see the Methods).

¹Department of Microbiology and Immunology, Dartmouth Medical School, Hanover, New Hampshire 03755, USA. [†]Present address: Department of Pathology and Laboratory Medicine, Hospital of the University of Pennsylvania, 3400 Spruce Street, 6 Founders, Philadelphia, Pennsylvania 19104, USA.

*These authors contributed equally to this work.

adherence to chitin and provide evidence that GbpA binds directly to GlcNAc residues in the context of chitin and presumably in the context of glycoproteins and lipids on intestinal epithelial cells.

To investigate further the affinity of GbpA for chitin, whole-cell lysates of $\Delta gbpA$, pGbpA-His *V. cholerae* were incubated with chitin beads (Methods). The GbpA-His fusion protein, expressed by *V. cholerae*, bound to the chitin beads (Fig. 2c) as judged by immunoblotting with a primary antibody directed against the histidine tag. Furthermore, this interaction was specific as judged by the fact that GbpA was the only protein detectable on a Coomassie-stained SDS polyacrylamide gel electrophoresis (SDS-PAGE) gel loaded with proteins that were attached to the chitin beads after incubation with whole-cell lysates.

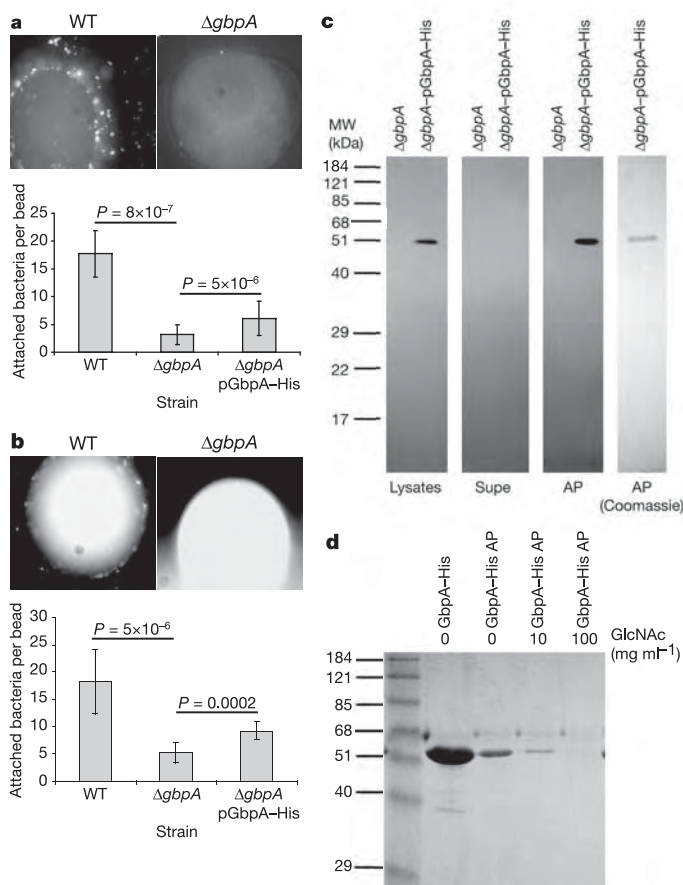


Figure 2 | GbpA is a GlcNAc-sensitive chitin-binding protein.

a, Micrographs of GFP expressing wild-type and $\Delta gbpA$ *V. cholerae* adhering to chitin beads, along with quantification (mean $\pm$ s.d.) of attached wild-type, $\Delta gbpA$ and $\Delta gbpA$, pGbpA-His *V. cholerae* to chitin beads.

b, Micrographs of GFP-expressing wild-type and $\Delta gbpA$ *V. cholerae* adhering to GlcNAc-coated beads, along with quantification (mean $\pm$ s.d.) of attached wild-type, $\Delta gbpA$ and $\Delta gbpA$, pGbpA-His *V. cholerae* to GlcNAc-coated beads. **c**, GbpA-His binds to chitin. A GbpA-His fusion protein was expressed in *V. cholerae* $\Delta gbpA$ and whole-cell lysates were prepared (lysates). The lysates were incubated with chitin beads for 1 h and adherent proteins were affinity purified. Proteins that remained bound to beads after three washes (AP) and proteins left in the lysate (supe) were separated by SDS-PAGE. Immunoblot analysis with antibodies directed against the histidine tag showed that GbpA binds to chitin, as judged by the presence of GbpA-His in the AP lanes. Coomassie staining further showed that, of the many proteins in the whole-cell lysates, GbpA-His was the most abundant protein associated with the chitin beads (last lane). **d**, Increasing concentrations of GlcNAc inhibit the attachment of purified GbpA-His to chitin beads. The GbpA-His lane contains the purified protein only and the AP lanes reflect the amount of GbpA-His bound to the chitin beads after washing.

To ensure that the $\Delta gbpA$ mutation was not affecting other bacterial surface proteins and thereby confounding these results, we verified the affinity of the pure GbpA protein for chitin and GlcNAc. Purified GbpA-His was mixed with chitin beads either in the absence or the presence of 10 or 100 mg ml⁻¹ of GlcNAc. GbpA-His attachment to the chitin beads was GlcNAc-sensitive, unambiguously showing that GbpA is a chitin-binding protein with specific affinity for GlcNAc (Fig. 2d). Reinforcing this principle, the addition of GlcNAc to our epithelial cell attachment assay reduced wild-type attachment to the cells (Supplementary Fig. S5).

Because the surfaces of some zooplankton are largely composed of chitin, we examined whether GbpA is important for *V. cholerae* attachment to zooplankton. *Daphnia magna* is a crustacean that sheds its chitinous exoskeleton on a regular basis. Previous studies have demonstrated a role for specific *V. cholerae* surface proteins in attachment to exoskeletons shed by the closely related organism, *Daphnia pulex*¹⁵. To investigate the role of GbpA in attachment to zooplankton *in vitro*, wild-type, $\Delta gbpA$ and $\Delta gbpA$, pGbpA-His *V. cholerae* (all bearing pGreen-Tir¹⁴) were incubated in buffer with *D. magna* exoskeletons. Attached bacteria were observed and counted by fluorescent microscopy. The mutant showed a nearly tenfold reduction, as compared with wild type, in attachment to the exoskeletons (Fig. 3a). The $\Delta gbpA$, pGbpA-His strain showed

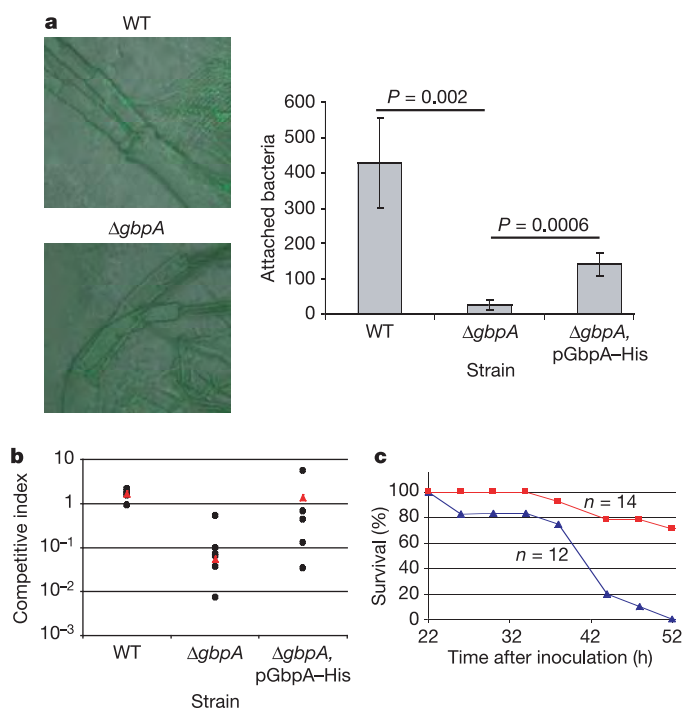


Figure 3 | GbpA is important for environmental and intestinal

colonization. **a**, Fluorescent micrograph of *D. magna* exoskeletons incubated with either wild-type or $\Delta gbpA$ *V. cholerae* bearing a plasmid expressing GFP. Quantification (mean $\pm$ s.d.) of wild type, $\Delta gbpA$ and $\Delta gbpA$, pGbpA-His attached to the *D. magna* exoskeletons is shown below the images. Nearly ten times more wild-type than *gbpA* mutant bacteria were attached. **b**, Competitive indices of wild-type, $\Delta gbpA$ and $\Delta gbpA$, pGbpA-His *V. cholerae* compared with a wild-type reference strain. Competitive index values for each mouse (circles) are shown along with the average for each group (triangle). The *gbpA* mutant was significantly attenuated for colonization, showing a roughly tenfold deficiency for colonization as compared with wild type. Expression of GbpA-His in the mutant strain partially complemented this phenotype. **c**, Survival curve showing the proportion of mice surviving up to 52 h after oral inoculation with 2.5×10^7 wild-type *V. cholerae* mixed with either preimmune (triangles) or hyperimmune (squares) serum from rabbits immunized with the GbpA-His protein.

increased binding as compared with $\Delta gbpA$, consistent with the partial complementation seen in the chitin and GlcNAc bead binding experiments. This result suggests a role for GbpA in bacterial attachment to, and colonization of, zooplankton in the aquatic ecosystem. This finding has implications for understanding the ecological interaction of *V. cholerae* in its natural habitat as well as for epidemiology. For example, it has been shown that removing large impurities (such as zooplankton) from drinking water in cholera endemic areas reduces disease incidence^{16,17}.

We considered that the significant deficiency for attachment of the $\Delta gbpA$ mutant to intestinal epithelial cells *in vitro* would correlate with a decrease in the ability of this mutant to colonize the small bowel *in vivo*, ultimately resulting in attenuated virulence. To test this, we first evaluated the ability of this mutant to colonize the infant mouse cholera model as compared with wild-type mice¹⁸. We assessed the impact of this mutation on virulence by determining the dose lethal to 50% of mice tested (LD_{50}) for this strain. The ratio of bacteria fed to the mice to bacteria collected from the mice the next day was, on average, tenfold reduced for the $\Delta gbpA$ strain as compared with wild type (Fig. 3b). When GbpA–His was expressed *in trans* in the $\Delta gbpA$ background, there was almost a complete restoration of the mutant's ability to colonize. Furthermore, the LD_{50} of wild type and that of $\Delta gbpA$, pGbpA–His *V. cholerae* were nearly identical, whereas that of the $\Delta gbpA$ strain was increased tenfold (data not shown). These results are significant in that they correlate the *in vitro* epithelial cell binding deficiency with intestinal colonization and virulence.

To demonstrate further a function for GbpA in intestinal colonization and to validate this protein as a potential vaccine target, passive immunization experiments were conducted with antisera from rabbits immunized with the GbpA–His protein. When mixed with wild-type bacteria before inoculation, the hyperimmune serum provided a significant survival advantage to the mice, thus further supporting a significant role for GbpA in colonization (Fig. 3c).

To localize the *gbpA* gene product with respect to *V. cholerae* subcellular compartments, whole-cell, periplasmic and culture supernatant fractions were probed for the presence of GbpA by immunoblotting (Fig. 4). Unexpectedly, the bulk of GbpA seemed to localize to the culture supernatant in wild-type *V. cholerae*, suggesting that GbpA is a secreted protein. Like many Gram-negative bacteria, *V. cholerae* possesses many specialized secretion systems designed to allow extracellular localization of proteins. The best-characterized secretion pathway elaborated by *V. cholerae* is the extracellular protein secretion (EPS) system. The most well-known substrate for this pathway is cholera toxin, although other substrates have been identified, including a chitinase¹⁹. To define a pathway for the secretion of GbpA, whole-cell extracts, periplasmic fractions and culture supernatants from an *epsD::kan* strain were probed for the presence of

GbpA²⁰. Notably, the *epsD::kan* mutant was significantly reduced for GbpA secretion as compared with wild type. These results support the conclusion that GbpA is a secreted protein that transits the bacterial outer membrane by means of the EPS secretion apparatus.

Our results indicate that GbpA functions to mediate attachment to both epithelial cell surfaces and chitin, most probably by a direct interaction with GlcNAc residues. This function enhances *V. cholerae* colonization in the gastrointestinal tract and may also be important in the environment by augmenting colonization of chitinous structures, leading to improved survival. It is important to consider that several adhesins are probably required for the initial attachment of *V. cholerae* to the intestinal epithelium and the loss of only one, as in the *gbpA* mutant, may not result in a complete attachment defect. The mechanism by which a secreted protein might mediate attachment is not obvious. It is possible that GbpA binds to GlcNAc after secretion and then interacts with a bacterial cell surface protein, forming a bridge between *V. cholerae* and the GlcNAc-coated surface. Alternatively, GbpA might be present in two forms, a secreted form and a (more minor) bacterial-surface-associated form. The fraction of GbpA that is associated with the bacterial surface may then mediate direct interactions between *V. cholerae* and GlcNAc-coated surfaces, or perhaps the two forms might interact with each other to promote attachment. Because much of GbpA is secreted, it is tantalizing to imagine that this secreted form may bind to GlcNAc surfaces alone, thereby reserving these surfaces for future *V. cholerae* attachment in the face of competing bacterial species. On the basis of these observations, it is almost certain that other virulence factors that have been studied in an attempt to understand human disease will also be found to have significant roles in the aquatic ecosystem that *V. cholerae* normally inhabits and that such parallels are likely to be widespread in the microbial world.

METHODS

Bacterial strains and plasmids. The *V. cholerae* classical O1 isolate CG842 was used as the wild-type strain in all experiments²¹. Transposon mutagenesis and selection of colonies for analysis were done as described²¹. The locations of the transposon insertion in mutants of interest were determined by chromosomal capture and DNA sequencing^{22,23}. The $\Delta gbpA$ mutant was constructed by allelic exchange as described²⁴. pGbpA–His was constructed by PCR amplification of the whole *gbpA* gene except for the stop codon from the *V. cholerae* chromosome and by cloning the resultant DNA into the expression vector pBAD–TOPO (Stratagene).

Cell culture, bacterial attachment assay and mutant screen. HT-29 cells were grown in DMEM medium supplemented with 10% fetal bovine serum and 1× Pen-Strep (Gibco) under 5% CO₂ as described²⁵. To screen for transposon mutants defective for epithelial attachment, roughly 400 candidate clones bearing random *TnphoA* insertions were each individually evaluated for their ability to attach to monolayers by this method. For epithelial attachment assays with defined mutations, each strain was tested in several wells and the results presented are the mean $\pm$ s.d. of all replicates.

Chitin and GlcNAc bead attachment assays. Chitin beads (New England Biolabs) or GlcNAc-coated agarose beads (Sigma) were washed three times in D-PBS buffer. Bacterial suspensions were added to the beads (10⁶ bacteria per ml) and the mixture was incubated at room temperature for 30 min. The beads were washed three times in 1 ml of D-PBS and prepared for fluorescent microscopy by mixing 50 μ l of the beads with an equal volume of Vectashield (Vector Labs) before coverslips were put in place. To analyse the ability of proteins to attach to the chitin beads, $\Delta gbpA$ and $\Delta gbpA$, pGbpA–His were grown overnight at 30 °C in LB medium (pH 6.5) containing 0.02% arabinose. The bacteria were collected by centrifugation and resuspended in PBS containing 20 μ M phenyl methylsulphonyl fluoride. Whole-cell lysates were prepared by sonicating bacterial suspensions (ten 10-s pulses). Lysates were cleared by centrifugation and 1 ml was added to 50 μ l of pre-washed (four 1-ml washes in PBS) chitin beads (New England Biolabs). The lysate–bead mixture was slowly rotated at 4 °C for 1 h. The chitin beads were collected by centrifugation, the supernatant was removed and reserved, and the bead fraction was washed five times in PBS. The chitin beads (along with attached proteins) were resuspended in 30 μ l of PBS plus an equal volume of 2× protein loading buffer and boiled for 5 min, and adherent proteins were separated on a 12.5% polyacrylamide gel containing SDS and visualized with Coomassie Brilliant Blue (Bio-Rad). For

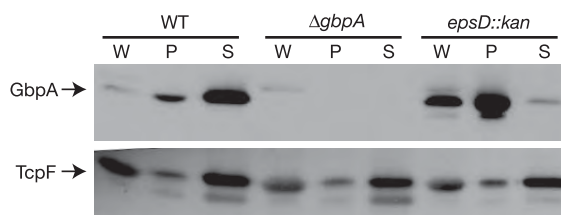


Figure 4 | GbpA is a secreted protein that requires EPS for extracellular localization. The relative abundance of GbpA and TcpF in whole-cell (W), periplasmic (P) and culture supernatant (S) fractions prepared from wild-type, $\Delta gbpA$ and *epsD::kan* *V. cholerae*, as determined by immunoblotting. GbpA was predominantly localized to the supernatant fractions in the wild-type strain but was reduced in the supernatant fraction prepared from the *epsD::kan* mutant. This is in contrast to TcpF, a protein known to be secreted independently of the EPS apparatus²⁶, which reached the supernatant in the wild-type, *epsD::kan* and $\Delta gbpA$ strains.

immunodetection, proteins were transferred to nitrocellulose at 4 °C in transfer buffer (25 mM Tris, 192 mM glycine and 20% methanol; pH 8.3) by using a wet transfer apparatus (Bio-Rad). Primary antibodies directed against the histidine tag were used at 1:1,000 (Qiagen) and secondary (horseradish peroxidase)-conjugated goat anti-mouse IgG (ICN) was used at 1:2,500.

D. magna exoskeleton attachment assays. We cultured *D. magna* in filtered pond water as described¹⁵. Cultures of *V. cholerae* expressing GFP were grown in LB for 18 h at 37 °C with aeration. Ten millilitres of KRT buffer (128 mM NaCl, 5.1 mM KCl, 1.34 mM MgSO₄, 2.7 mM CaCl₂ and 10 mM Tris-HCl; pH 7.5) containing three or more *D. magna* exoskeletons was inoculated with *V. cholerae* from overnight cultures at a 1:100 dilution and the mixture was incubated at 23 °C for 2 h. Exoskeletons were then rinsed serially three times in 20 ml of KRT buffer by gentle stirring and transfer with a pipette. Exoskeletons were slide-mounted with Vectashield (Vector Laboratories) and fluorescent images were digitally captured using OPENLAB software.

Infant mouse cholera model. Five-day-old CD-1 mice from mixed litters were orally inoculated with 50 µl of a 1×10^{-2} dilution of overnight bacterial culture (grown in LB at 30 °C with a starting pH of 6.5) and incubated at 30 °C for 24 h. The bacteria were then recovered by homogenizing mouse intestines in 4 ml of a 10% glycerol solution in LB. The homogenate was appropriately diluted and plated on solid medium containing streptomycin. The competitive index was calculated by comparing the ratio of mutant bacteria recovered from the intestine to mutant strain input to the input/output ratio of the wild-type strain in matched control mice inoculated with wild-type bacteria. The LD₅₀ of each strain was assessed as described²⁶.

For antisera protection assays, sera from rabbits before and after immunization with the purified GbpA-His protein were obtained from Pocono Rabbit Farm. The sera were mixed at a 1:1 ratio (v/v) with a 1:100 dilution of wild-type *V. cholerae* overnight culture. Mice were inoculated as above and the number of live and dead mice was recorded every 4 h.

Bacterial fractionation. Fractionation of proteins was done as described²⁶. To localize GbpA, primary antisera directed against GbpA-His were used at 1:1,000 dilution and a secondary goat antiserum against rabbit IgG was used at 1:100,000 dilution (Cappel). The blot was stripped and TcpF was detected with polyclonal antisera at 1:1,000 dilution²⁷.

Received 20 July; accepted 19 September 2005.

- Reidl, J. & Klose, K. E. *Vibrio cholerae* and cholera: out of the water and into the host. *FEMS Microbiol. Rev.* **26**, 125–139 (2002).
- Tamplin, M. L., Gauzens, A. L., Huq, A., Sack, D. A. & Colwell, R. R. Attachment of *Vibrio cholerae* serogroup O1 to zooplankton and phytoplankton of Bangladesh waters. *Appl. Environ. Microbiol.* **56**, 1977–1980 (1990).
- Cottingham, K. L., Chiavelli, D. A. & Taylor, R. K. Environmental microbe and human pathogen: the ecology and microbiology of *Vibrio cholerae*. *Front. Ecol. Environ.* **1**, 80–86 (2003).
- Huq, A., Small, E., West, P. & Colwell, R. R. The role of planktonic copepods in the survival and multiplication of *Vibrio cholerae* in the aquatic environment. in *Vibrios in the Environment* (ed. Colwell, R. R.) 521–534 (John Wiley & Sons, New York, NY, 1984).
- Kaper, J. B., Morris, J. G. & Levine, M. M. Cholera. *Clin. Microbiol. Rev.* **8**, 48–86 (1995).
- Heidelberg, J. F. et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–483 (2000).
- Nalin, D. R., Daya, V., Reid, A., Levine, M. M. & Cisneros, L. Adsorption and growth of *Vibrio cholerae* on chitin. *Infect. Immun.* **25**, 768–770 (1979).
- Meibom, K. L. et al. The *Vibrio cholerae* chitin utilization program. *Proc. Natl Acad. Sci. USA* **101**, 2524–2529 (2004).
- Zampini, M. et al. *Vibrio cholerae* persistence in aquatic environments and colonization of intestinal cells: involvement of a common adhesion mechanism. *FEMS Microbiol. Lett.* **244**, 267–273 (2005).
- Tarsi, R. & Pruzzo, C. Role of surface proteins in *Vibrio cholerae* attachment to chitin. *Appl. Environ. Microbiol.* **65**, 1348–1351 (1999).
- Sasmal, D., Guhathakurta, B., Ghosh, A. N., Pal, C. R. & Datta, A. N-acetyl-D-glucosamine-specific lectin purified from *Vibrio cholerae* O1. *FEMS Microbiol. Lett.* **77**, 217–224 (1992).
- Finne, J. et al. Novel polyfucosylated N-linked glycopeptides with blood group A, H, X, and Y determinants from human small intestinal epithelial cells. *J. Biol. Chem.* **264**, 5720–5735 (1989).
- Bjork, S., Breimer, M. E., Hansson, G. C., Karlsson, K. A. & Leffler, H. Structures of blood group glycosphingolipids of human small intestine. A relation between the expression of fucolipids of epithelial cells and the ABO, Le and Se phenotype of the donor. *J. Biol. Chem.* **262**, 6758–6765 (1987).
- Miller, W. G. & Lindow, S. E. An improved GFP cloning cassette designed for prokaryotic transcriptional fusions. *Gene* **191**, 149–153 (1997).
- Chiavelli, D. A., Marsh, J. W. & Taylor, R. K. The mannose-sensitive hemagglutinin of *Vibrio cholerae* promotes adherence to zooplankton. *Appl. Environ. Microbiol.* **67**, 3220–3225 (2001).
- Huo, A. et al. A simple filtration method to remove plankton-associated *Vibrio cholerae* in raw water supplies in developing countries. *Appl. Environ. Microbiol.* **62**, 2508–2512 (1996).
- Colwell, R. R. et al. Reduction of cholera in Bangladeshi villages by simple filtration. *Proc. Natl Acad. Sci. USA* **100**, 1051–1055 (2003).
- Sun, D. X., Mekalanos, J. J. & Taylor, R. K. Antibodies directed against the toxin-coregulated pilus isolated from *Vibrio cholerae* provide protection in the infant mouse experimental cholera model. *J. Infect. Dis.* **161**, 1231–1236 (1990).
- Connell, T. D., Metzger, D. J., Lynch, J. & Folster, J. P. Endochitinase is transported to the extracellular milieu by the eps-encoded general secretory pathway of *Vibrio cholerae*. *J. Bacteriol.* **180**, 5591–5600 (1998).
- Davis, B. M. et al. Convergence of the secretory pathways for cholera toxin and the filamentous phage, CTXφ. *Science* **288**, 333–335 (2000).
- Taylor, R. K., Miller, V. L., Furlong, D. B. & Mekalanos, J. J. Use of phoA gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proc. Natl Acad. Sci. USA* **84**, 2833–2837 (1987).
- Skorupski, K. & Taylor, R. K. Sequence and functional analysis of the gene encoding *Vibrio cholerae* cAMP receptor protein. *Gene* **198**, 297–303 (1997).
- Skorupski, K. & Taylor, R. K. Cyclic AMP and its receptor protein negatively regulate the coordinate expression of cholera toxin and toxin-coregulated pilus in *Vibrio cholerae*. *Proc. Natl Acad. Sci. USA* **94**, 265–270 (1997).
- Skorupski, K. & Taylor, R. K. Positive selection vectors for allelic exchange. *Gene* **169**, 47–52 (1996).
- Benitez, J. A. et al. Adherence of *Vibrio cholerae* to cultured differentiated human intestinal cells: an in vitro colonization model. *Infect. Immun.* **65**, 3474–3477 (1997).
- Kirn, T. J. & Taylor, R. K. TcpF is a soluble colonization factor and protective antigen secreted by El Tor and classical O1 and O139 *Vibrio cholerae* serogroups. *Infect. Immun.* **73**, 4461–4470 (2005).
- Kirn, T. J., Bose, N. & Taylor, R. K. Secretion of a soluble colonization factor by the TCP type 4 pilus biogenesis pathway in *Vibrio cholerae*. *Mol. Microbiol.* **49**, 81–92 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Sandoe for technical assistance. This work was supported by the NIH and a Rosalind Borison memorial fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.K.T. (Ronald.K.Taylor@dartmouth.edu).

Casein kinase 1 γ couples Wnt receptor activation to cytoplasmic signal transduction

Gary Davidson^{1*}, Wei Wu^{1*}, Jinlong Shen¹, Josipa Bilic¹, Ursula Fenger¹, Peter Stannek¹, Andrei Glinka¹ & Christof Niehrs¹

Signalling by Wnt proteins (Wingless in *Drosophila*) has diverse roles during embryonic development and in adults, and is implicated in human diseases, including cancer^{1,2}. LDL-receptor-related proteins 5 and 6 (LRP5 and LRP6; Arrow in *Drosophila*) are key receptors required for transmission of Wnt/ β -catenin signalling in metazoa³. Although the role of these receptors in Wnt signalling is well established, their coupling with the cytoplasmic signalling apparatus remains poorly defined. Using a protein modification screen for regulators of LRP6, we describe the identification of *Xenopus* Casein kinase 1 γ (CK1 γ), a membrane-bound member of the CK1 family. Gain-of-function and loss-of-function experiments show that CK1 γ is both necessary and sufficient to transduce LRP6 signalling in vertebrates and *Drosophila* cells. In *Xenopus* embryos, CK1 γ is required during antero-posterior patterning to promote posteriorizing Wnt/ β -catenin signalling. CK1 γ is associated with LRP6, which has multiple, modular CK1 phosphorylation sites. Wnt treatment induces the rapid CK1 γ -mediated phosphorylation of these sites within LRP6, which, in turn, promotes the recruitment of the scaffold protein Axin. Our results reveal an evolutionarily conserved mechanism that couples Wnt receptor activation to the cytoplasmic signal transduction apparatus.

In a human cell-culture-based small pool expression screen for proteins able to covalently modify LRP6, we identified CK1 γ , a protein which retards ('upshifts') the migration of LRP6 on SDS-polyacrylamide gel electrophoresis (SDS-PAGE) in co-transfection assays (Supplementary Fig. 1a; Fig. 1a). There are three closely related isoforms of CK1 γ —CK1 γ 1, 2 and 3—that are unique within the CK1 family in carrying putative palmitoylation sites at the carboxy terminus (TKCCFFFR; Supplementary Fig. 2)⁴, which anchor the proteins in the plasma membrane (see below). This is unlike CK1 α , δ and ϵ , which are cytoplasmic or nuclear and are involved in the direct regulation of Wnt pathway components other than LRP5 and LRP6 (ref. 5). Indeed, LRP6 upshift is promoted by CK1 γ , while CK1 ϵ has no effect on the electrophoretic mobility of LRP6 (Fig. 1a). The cytoplasmic domain of LRP6 is the region modified by CK1 γ , as an LRP6- Δ N protein that lacks the amino-terminal extracellular domain (ECD) is still able to produce the higher molecular weight bands indicative of CK1 γ -mediated modification, whereas an LRP6- Δ C protein that lacks the carboxy-terminal intracellular domain (ICD) does not (Fig. 1a). These results suggest that the C terminus of LRP6 is modified by CK1 γ phosphorylation, and this was confirmed by ³²P-incorporation *in vivo* (Fig. 1b). Moreover, modification of LRP6 by CK1 γ enhances LRP6 function in Wnt reporter assays, in which LRP6 synergizes strongly with CK1 γ and more weakly with CK1 ϵ , while Dishevelled-1 (Dvl1) shows the inverse behaviour (Fig. 1c), consistent with the latter being directly regulated by

CK1 ϵ (ref. 6). Similar gain-of-function results were obtained in *Xenopus* embryos (Supplementary Fig. 1b).

To investigate the requirement of CK1 γ in LRP6 signalling, we designed two dominant-negative forms of CK1 γ with point mutations (K73R or D164N) in the ATPase domain, previously identified as generating specific inhibitors for CK1 ϵ (ref. 6). Both dominant-negative forms of CK1 γ inhibit the upshift, and hence phosphorylation, of LRP6 by wild-type CK1 γ , but phosphorylation of Dvl1 by CK1 ϵ (ref. 6) is unaffected (Supplementary Fig. 1d). Therefore, both dominant-negative forms of CK1 γ specifically inhibit CK1 γ . Dominant-negative CK1 γ inhibits Wnt signalling in HEK293T cells (Fig. 1d), and similar loss-of-function results were obtained in *Xenopus* embryos (Supplementary Fig. 1e–g); wild-type CK1 γ rescues this effect in both cases (Supplementary Fig. 1f and data not shown). Unlike Wnt1, Wnt3a and a constitutively active LRP6- Δ E(1–4) protein, lacking most of the extracellular domain⁷, β -catenin activity is unaffected by dominant-negative forms of CK1 γ (Fig. 1d). This is consistent with CK1 γ acting directly on LRP6.

Wingless (Wg) reporter assays in *Drosophila* SL2 cells show that the single *Drosophila* homologue of CK1 γ , Gilgamesh (Gish)⁸, strongly synergizes with the *Drosophila* LRP6 homologue, Arrow (Supplementary Fig. 1c). Conversely, using an interfering RNA (RNAi) loss-of-function approach, a *gish* double-stranded RNA that targets all eight *gish* transcripts blocks Wg but not Dishevelled (Dsh) signalling (Fig. 1e, left panel versus right panel). This is unlike *armadillo* (*arm*) and *dsh* dsRNAs, which block signalling by both Wg and Dsh (Fig. 1e). These results support an evolutionarily conserved role for CK1 γ in LRP6 regulation. The absence of obvious Wg phenotypes in *Drosophila gish* mutants¹ may be because at least two alternatively spliced transcripts within the gene locus are intact⁸.

In *Xenopus* embryos, maternally derived *ck1 γ* mRNA transcripts are present ubiquitously at gastrulation, and zygotic transcription starts at around the mid-neurula stage (Supplementary Fig. 3). To address the physiological role of CK1 γ , we performed *Xenopus* gain-of-function and loss-of-function experiments (Fig. 1f, g; Supplementary Fig. 1g; Supplementary Fig. 4). Overexpression of CK1 γ in the animal pole by messenger RNA injection leads to headless embryos (Fig. 1f), whereas injection of either a *ck1 γ* morpholino or dominant-negative *ck1 γ ^{K73R}* mRNA reduces trunk and tail structures, and induces enlarged heads and cement glands (Fig. 1g; Supplementary Fig. 4b–d, h, i). These phenotypes match hyper- and hypo-activation of zygotic Wnt signalling, respectively. We conclude that CK1 γ is required for Wnt-mediated antagonism of Spemann's head organizer.

Following transfection in HEK293T cells, CK1 γ tagged with enhanced yellow fluorescent protein (EYFP-CK1 γ) localizes to the plasma membrane, consistent with it containing a C-terminal

¹Division of Molecular Embryology, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.

*These authors contributed equally to this work.

palmitoylation site. Deleting this region (CK1 γ - Δ C) results in the cytoplasmic localization of CK1 γ (Fig. 2a), and abolishes Wnt/ β -catenin signalling (Fig. 2b), although its kinase activity is unaffected (Fig. 2c). Bioluminescence resonance energy transfer (BRET) assays⁹ reveal that CK1 γ and LRP6 interact in live cells, and that this interaction requires membrane localization, as it is abolished in CK1 γ - Δ C (Fig. 2d). Co-immunoprecipitation (CoIP) confirms that LRP6 specifically interacts with CK1 γ and not CK1 γ - Δ C (data not shown), and this is independent of coexpressed Axin or GSK3 β (Fig. 2e). Neither BRET nor CoIP assays show differences in CK1 γ -LRP6 interaction on Wnt treatment (data not shown). These results indicate that CK1 γ and LRP6 associate, and that this association requires the membrane localization of CK1 γ .

To identify relevant phosphorylation sites in LRP6, serial C-terminal deletions were tested for loss of CK1 γ synergy in Wnt reporter assays. The LRP6- Δ E(1–4) backbone generates a robust Wnt response (Fig. 3a). A gradual, not abrupt, decrease in reporter activation with progressive deletions is observed, consistent with the presence of multiple activating phosphorylation sites. LRP6- Δ E(1–4)- Δ 87, which retains a single PPPSP motif, was identified as the shortest construct still activated by CK1 γ (Fig. 3a).

Deletion of Ser and/or Thr clusters in LRP6- Δ E(1–4)- Δ 87 revealed that two regions are required for cooperation with CK1 γ (Supplementary Fig. 5, underlined sequences numbered 6 and 7). To test if

the two regions are also sufficient for Wnt signalling, we linked them alone, or in combination, to LDLR- Δ N, a heterologous low-density lipoprotein (LDL) mini-receptor structurally analogous to LRP6- Δ N but inactive in Wnt signalling¹⁰. Neither region alone (miniA or miniB) activates Wnt signalling, but they do when combined (miniC), and this effect is further enhanced by CK1 γ both in HEK293T cells (Fig. 3b) and *Xenopus* embryos (data not shown).

The region of LRP6 contained within miniC consists of a PPPSP motif flanked by two evolutionarily conserved Ser/Thr clusters that are potential CK1 phosphorylation sites (see the underlined sequences in Fig. 3d and Supplementary Fig. 6). We refer to the N-terminal CK1 site (TGA(S)₇TKGT) as cluster 1, and the C-terminal CK1 site (SPATERSHYT) as cluster 2 (abbreviated 1/S/2, with the Ser residue in the PPPSP motif designated by the S). To analyse the function of the two CK1 site clusters in Wnt signalling, we tested inactivating Ala (A) and phospho-mimicking Asp (D) substitutions of all Ser/Thr residues within these clusters (shown above in bold), using the miniC receptor (see Fig. 3b) as a backbone. The wild-type mini-receptor (1/S/2) is constitutively active, unlike A/S/A, indicating that the CK1 sites are required for Wnt signalling (Fig. 3c). Elimination of either CK1 site alone (A/S/2 or 1/S/A) suppresses Wnt signalling. Conversely, Asp substitutions in individual (D/S/2 or 1/S/D) or both CK1 sites (D/D/2) lead to Wnt hyperactivation. Constitutive Wnt signalling by 1/S/2, D/S/2 and 1/S/D is blocked by

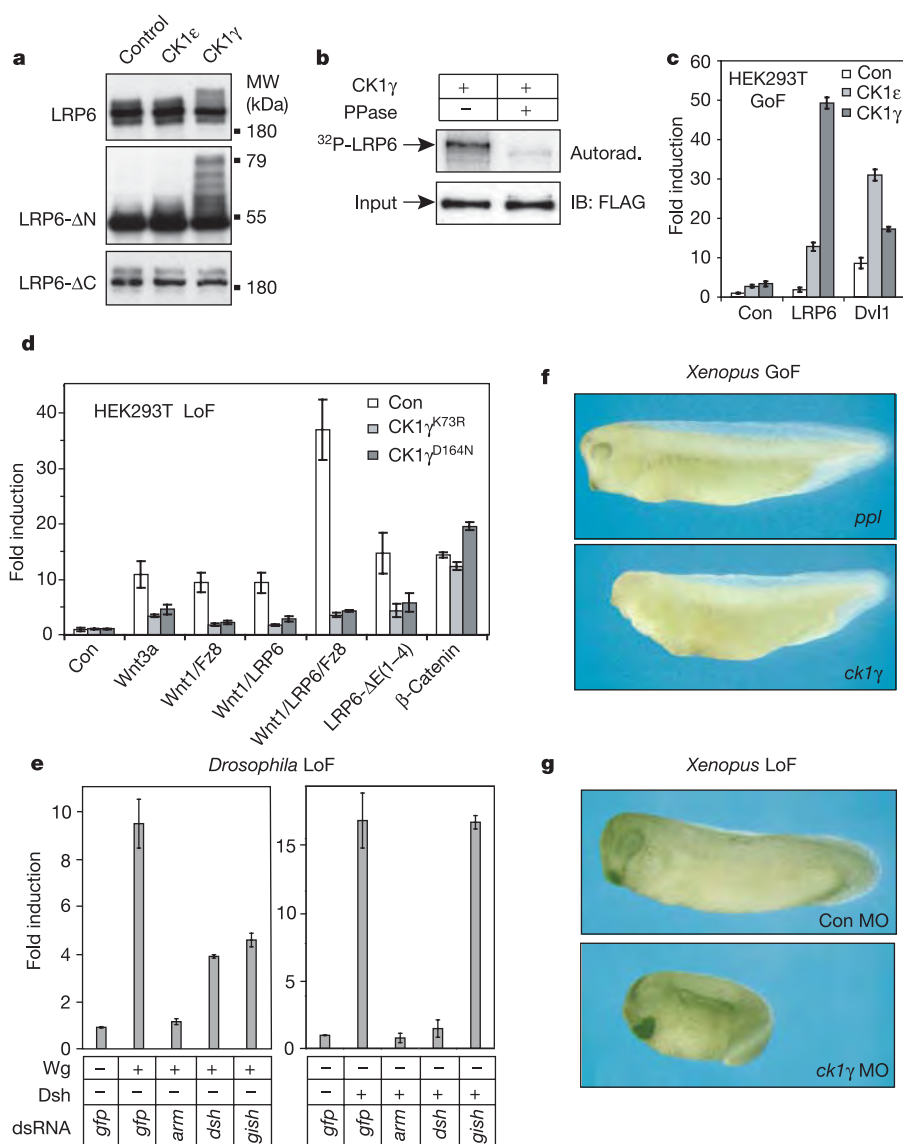


Figure 1 | Identification of CK1 γ as a specific modifier of LRP6.

a, The cytoplasmic domain of LRP6 is modified by CK1 γ . Anti-Myc immunoblots of wild-type Myc-LRP6 or LRP6 deletions coexpressed with CK1 isoforms in HEK293T cells as indicated. MW, molecular weight. **b**, ³²P-labelling of LRP6 *in vivo*. Autorad., autoradiography of ³²P-incorporation; IB, immunoblot using an anti-FLAG antibody.

c, d, Wnt luciferase reporter assays in HEK293T cells showing the sufficiency (**c**) and requirement (**d**) of CK1 γ in Wnt/LRP signalling. Error bars show s.d. GoF, gain of function; LoF, loss of function; Con, control empty vector. **e**, Wg luciferase reporter assays in SL2 cells showing requirement of the single *Drosophila* CK1 γ homologue, Gish, in Wg signalling. **f, g**, *Xenopus* embryos microinjected at the animal pole with either 500 pg human *ck1γ1* mRNA (**f**) or 60 ng *ck1γ1* antisense morpholino oligonucleotide (**g**) are posteriorized and anteriorized, respectively (see also Supplementary Fig. 4). *ppl*, preprolactin; MO, morpholino.

dominant-negative CK1 γ , indicating that both clusters require CK1 γ for Wnt signalling. D/S/D, on the other hand, in which both CK1 sites are artificially activated, signals independently of CK1 γ . Any substitution of the PPPSP site (1/A/2 or 1/D/2) inactivates Wnt signalling, indicating that this serine is essential. The results suggest that (1) clusters 1 and 2 are both physiologically relevant CK1 γ phosphorylation sites that activate Wnt signalling, and (2) they require an intact PPPSP motif.

To show direct phosphorylation of LRP6 by CK1 γ , we performed *in vitro* kinase assays. We raised a phospho-specific antibody (Tp1479) against a CK1 site within cluster 1, which recognizes the most C-terminal phosphorylated threonine (Fig. 3d, bottom panel; Supplementary Fig. 6, 7). *In vitro* kinase assays were performed with CK1 γ using the miniC receptor carrying various internal mutations as the substrate (Fig. 3c). Following the kinase reaction and SDS-PAGE, samples were analysed for 32 P-incorporation by autoradiography (Fig. 3d, top panel), and T1479 phosphorylation by immunoblotting (Fig. 3d, bottom panel). The wild-type mini-receptor (1/S/2) is phosphorylated by CK1 γ , unlike the construct in which all three sites are inactivated (A/A/A). In agreement with cluster 1 being a CK1 phosphorylation site, 1/A/A is robustly phosphorylated. In contrast, A/S/A, containing only the intact PPPSP site, is not phosphorylated by CK1 γ . Notably, cluster 2 is only weakly phosphorylated (A/A/2), but this is enhanced either by the presence of an intact PPPSP motif (A/S/2) or mimicking cluster 1 and PPPSP phosphorylation (D/D/2). Taken together, these results indicate that clusters 1 and 2, but not the PPPSP motif, are direct target sites for CK1 γ phosphorylation, and that efficient phosphorylation of cluster 2 requires N-terminal priming from the PPPSP site.

T1479 is phosphorylated in all constructs containing an unmodified cluster 1 (Fig. 3d, bottom panel), indicating that it is a *bona fide* target for phosphorylation by CK1 γ .

An important consequence of Wnt activation is the sequestration of Axin, a negative regulator of β -catenin^{11,12}, by LRP6. Therefore, we asked whether CK1 γ is required for Axin recruitment. In CoIP assays, dominant-negative CK1 γ markedly reduces binding of Axin to LRP6 (Fig. 4a). It has been shown that the PPPSP site of LRP6 is required for Axin binding¹⁰. To determine if the two CK1 phosphorylation site clusters in LRP6 also play a role in Axin binding, CoIPs were performed using 1/S/2 site substitutions within the LRP6- Δ E(1-4)- Δ 87 protein (Fig. 4b). While only weak interaction is seen with the wild-type construct (1/S/2), mimicking CK1 γ phosphorylation of cluster 1 in D/S/2 leads to robust binding of Axin. Increased binding is also observed in 1/S/D, supporting the notion that phosphorylation of cluster 2 (Fig. 3d) is physiologically relevant. However, activation of this cluster by itself (A/S/D) fails to mediate Axin binding. Likewise, D/S/2 shows stronger binding than D/S/A, arguing that both CK1 sites act synergistically to promote Axin binding when they are phosphorylated. Of note, 1/D/D fails to bind Axin, indicating that a phospho-serine at the PPPSP site cannot be mimicked by an Asp, explaining the inactivity of 1/D/2 in Fig. 3c. We conclude that CK1 γ is required for Axin binding, and that both of the identified CK1 phosphorylation site clusters promote this interaction.

Wnt signalling promotes Axin binding to LRP6 (ref. 11), and our data indicate that this is mediated by CK1 γ phosphorylation at conserved sites in LRP6. An important prediction, therefore, is that Wnt stimulation leads to phosphorylation of LRP6 at sites of CK1 γ

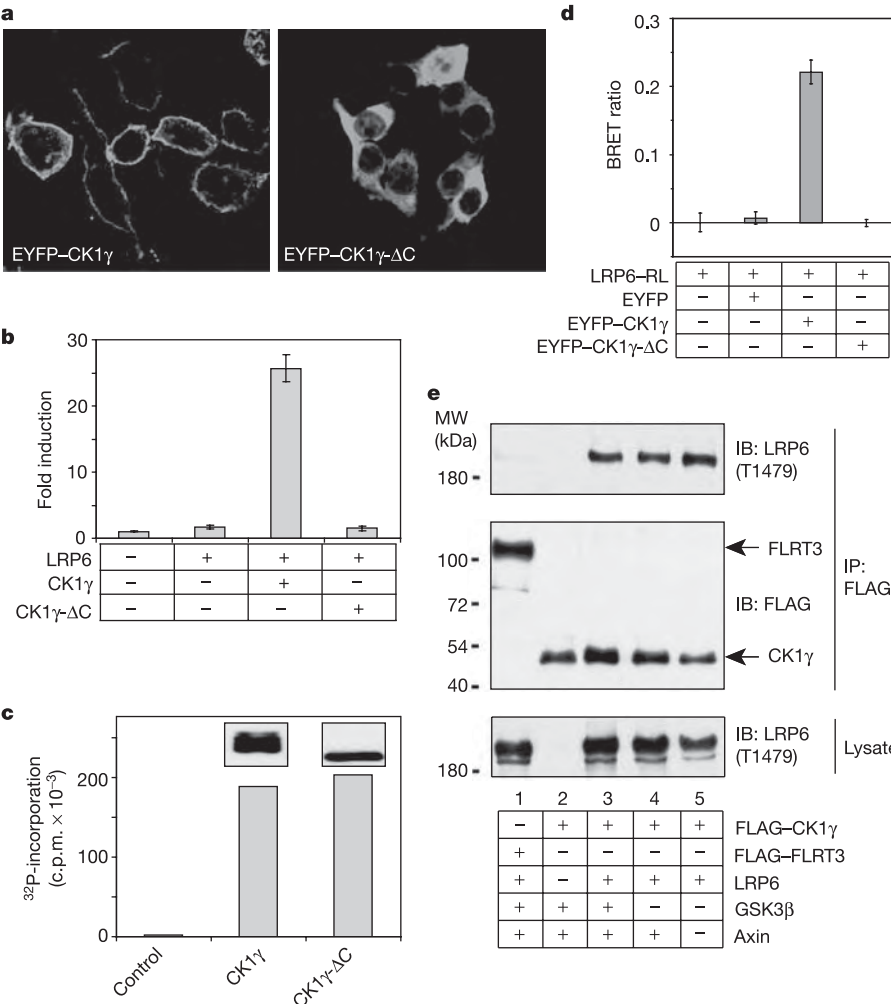


Figure 2 | CK1 γ localizes at the cell membrane and associates with LRP6. **a**, Confocal images of indicated proteins in transfected HEK293T cells. **b**, Wnt luciferase reporter assay in HEK293T cells. Error bars show s.d. **c**, *In vitro* kinase assay using a CK1 peptide substrate and the indicated CK1 γ proteins immunoprecipitated from transfected HEK293T cells. Inset shows CK1 γ protein production controls. **d**, BRET assay in live HEK293T cells expressing the indicated proteins. See the Methods for details of the BRET assay. RL, *Renilla* luciferase. **e**, Immunoblots (IB) of immunoprecipitates (IP) from cells expressing the indicated proteins. FLRT3, Fibronectin leucine-rich transmembrane protein 3.

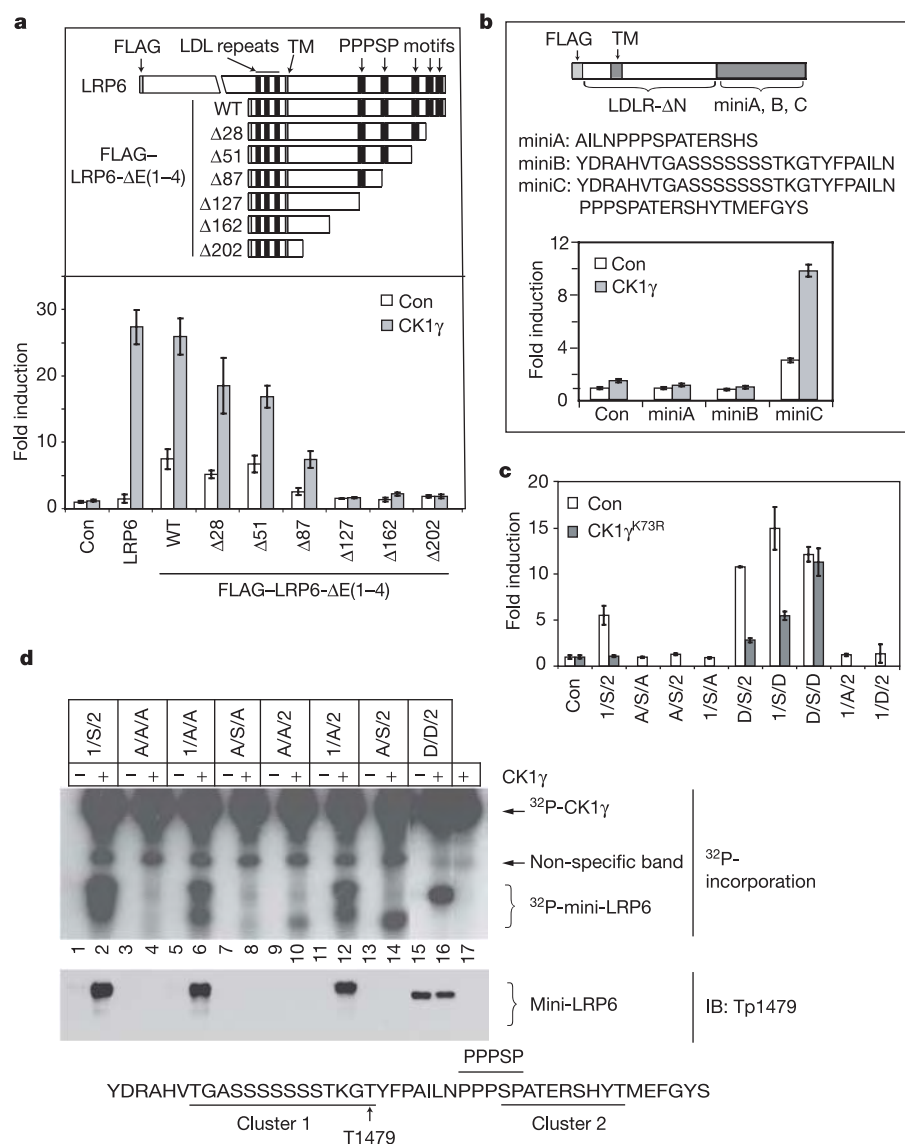


Figure 3 | CK1 γ phosphorylates two Ser/Thr clusters flanking a PPPSP motif in LRP6. **a**, Wnt luciferase reporter assay in HEK293T cells with serial LRP6 C-terminal deletions identifying a minimal region for CK1 γ activity. Schematic representation of full-length and C-terminal-deleted FLAG-LRP6 proteins. Error bars show s.d. Con, control empty vector; TM, transmembrane domain; WT, wild type. **b**, Wnt luciferase reporter assay in HEK293T cells with LDLR-ΔN mini-receptors. Schematic representation of the mini-receptors and amino acid sequences of the linked motifs from LRP6 are shown above. **c**, Wnt luciferase reporter assay in HEK293T cells with mini-receptors, with or without co-transfected dominant-negative *ck1γ*. MiniC is used as the wild-type mini-receptor (1/S/2), and all other derivatives are mutants of this construct. **d**, CK1 γ *in vitro* kinase assay using mini-receptors as substrates. Autoradiography (top panel) or anti-Tp1479 immunoblot (bottom panel) shows ³²P-incorporation and CK1 γ -phosphorylated T1479, respectively. Anti-FLAG immunoblots showed that all proteins were similarly produced (data not shown). The linked LRP6 region within the wild-type miniC receptor is shown beneath, with CK1 γ clusters 1 and 2, as well as the PPPSP site, underlined. The arrow indicates T1479 against which a phospho-specific antibody was raised. Constructs in which cluster 1 has been mutated to A are not recognized by this antibody.

phosphorylation. Indeed, CK1 γ overexpression in HEK293T cells induces phosphorylation at T1479 (in cluster 1), and this can be mimicked by Wnt treatment (Fig. 4c). This Wnt-induced T1479 phosphorylation requires CK1 γ , as it is blocked by the dominant-negative mutant CK1 γ ^{K73R}. The PPPSP motifs are also thought to become phosphorylated following Wnt stimulation¹⁰. However, using a phospho-specific anti-PPSP antibody (Sp1490; Supplementary Fig. 6, 7), no significant phosphorylation increase is detected after either CK1 γ overexpression or Wnt treatment (Fig. 4c, middle panel). Strikingly, as little as 10 min of Wnt stimulation is sufficient to induce robust T1479 phosphorylation of endogenous LRP6 in mouse embryonic fibroblasts (MEFs; Fig. 4d, top panel), P19 and HeLa cells (data not shown). Unlike T1479, S1490 (in the PPPSP motif) is constitutively phosphorylated in all cell lines tested, and shows either no (data not shown) or weaker Wnt induction (Fig. 4d, middle panel). Using a phospho-independent T1479 antibody, which recognizes total LRP6, a progressive mobility upshift is observed, which saturates after 60 min of Wnt treatment (Fig. 4d, bottom panel). We conclude that Wnt stimulation results in phosphorylation of T1479 (cluster1) in a CK1 γ -dependent manner.

We and others suggested that the ECD of LRP6 exerts an auto-inhibitory effect that is relieved upon Wnt stimulation, based on the observation that its deletion leads to constitutive receptor activation^{7,11,13}. Alternatively, it was suggested that deletion of the ECD

enhances transport to the plasma membrane, because co-transfection of the chaperone MESD (for mesoderm development) with full-length LRP6 activates Wnt signalling as effectively as the ECD-deleted form¹⁴. To distinguish between these two possibilities we coexpressed either full-length LRP6 and MESD or LRP6-ΔE(1-4) alone in HEK293T cells and analysed T1479 phosphorylation. No signal is detected in full-length LRP6 despite overexpressing MESD (Fig. 4e), while there is a strong signal in LRP6-ΔE(1-4). A slight increase in S1490 (PPPSP) phosphorylation was also observed in LRP6-ΔE(1-4) compared to full-length LRP6. Our results therefore support the idea of an autoinhibitory role of the LRP6 ECD, and indicate that one of its functions is to prevent phosphorylation by CK1 γ .

In summary, this investigation indicates that CK1 γ couples the extracellular Wnt signal and the cytoplasmic signal transduction machinery, and suggests a model where the net effect of CK1 γ phosphorylation of LRP6 is Axin recruitment (Supplementary Fig. 8). The cytoplasmic domain of LRP6 contains five reiterated PPPSP motifs that are necessary for Wnt/ β -catenin signalling. While the PPPSP motifs are phosphorylated, our data demonstrate that this is not by CK1 γ , but by an unknown, probably proline-directed, kinase. As Wnt was reported to stimulate PPPSP phosphorylation¹⁰, we were surprised to find not only that there is constitutive S1490 (PPPSP) phosphorylation in unstimulated cells, but that Wnt stimulates no or only weak induction at this site. Since CK1 members

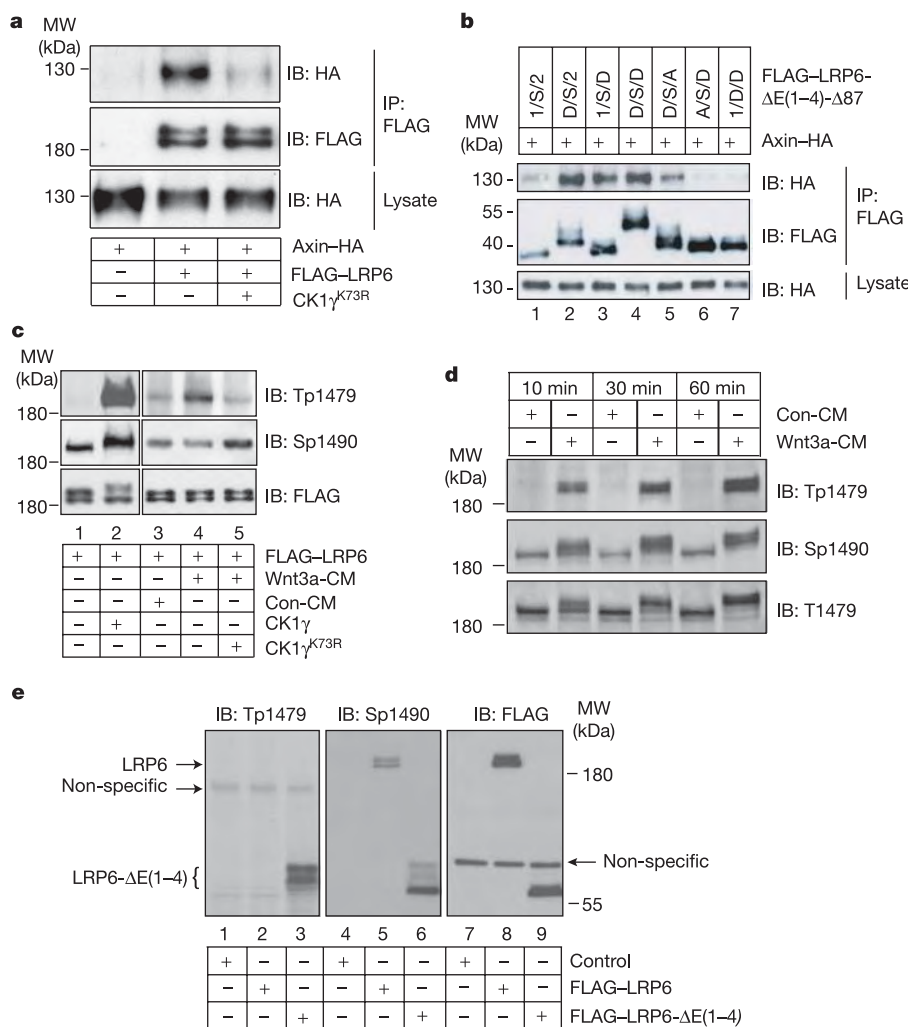


Figure 4 | CK1 γ mediates Axin recruitment and Wnt-induced phosphorylation at threonine 1479. **a, b**, Immunoblots (IB) from CoIPs (IP) of HEK293T cells expressing the indicated proteins. **b**, FLAG-LRP6- Δ E(1-4)- Δ 87 (1/S/2) and its corresponding mutants were coexpressed with Axin-HA. Amino acid substitutions were as in Fig. 3d. **c**, Immunoblots of HEK293T cells expressing the indicated proteins and stimulated for 2 h with control conditioned medium (Co-CM) or Wnt3a conditioned medium (Wnt3a-CM). **d**, Immunoblots of MEFs stimulated with control conditioned medium (Co-CM) or Wnt3a conditioned medium (Wnt3a-CM) for the indicated times. **e**, Immunoblots of HEK293T cells expressing the indicated proteins.

act on proteins previously phosphorylated by other kinases, constitutive S1490 (PPPSP) phosphorylation may prime CK1 γ -mediated phosphorylation. This may well apply to all PPPSP sites in LRP6, as all are succeeded by at least one adjacent CK1 site (Supplementary Fig. 6). We therefore propose that the LRP6 cytoplasmic domain is modular, and each element consists of a CK1 γ phosphorylation site that is primed by a phosphorylated PPPSP motif.

METHODS

LRP6 protein modification screen and constructs. A *Xenopus tropicalis* complementary DNA library was prepared¹⁵ with mRNA from embryos at stage 10, 20 and 30. Individual colonies (100,000) were arrayed in 384-well plates and, from each plate, four sets of 96 colonies were combined to prepare plasmid DNA pools. DNA (200 ng) from each pool was co-transfected with 10 ng 6 $\times$ Myc-LRP6 plasmid (encoding Myc-tagged human LRP6) into HEK293T cells in 96-well plates. After transfection (40 h), cells were lysed in 2% NP40 buffer, proteins separated by 7–12% SDS-PAGE, and anti-Myc immunoblots carried out. Single clones from positive pools were isolated after secondary screening of the original 384-well plate. V5-*krm2* and V5-*dkk3* were then co-transfected as controls. CK1 γ ^{K73R} and CK1 γ ^{D164N} were made by substitution of lysine (K) 73 with arginine (R) and of aspartic acid (D) 164 with asparagine (N), respectively, in *Xenopus* CK1 γ 1. CK1 γ - Δ C was made by deletion of the last 20 amino acids of *Xenopus* CK1 γ 1. LDLR- Δ N mini-receptors were made as described¹⁰. All deletion and mutation plasmids were constructed by polymerase chain reaction (PCR).

Cell culture, luciferase reporter and BRET assays. HEK293T, MEF, P19 and HeLa cell lines were maintained in DMEM, 10% FCS and 10% CO₂. Mouse Wnt3a conditioned medium (Wnt3a-CM) was produced from mouse L cells stably transfected with mouse *Wnt3a* (American Type Culture Collection CRL-2647) and control conditioned medium (Co-CM) was from non-transfected L cells¹⁶.

Luciferase reporter assays in HEK293T cells were carried out in 96-well plates as described¹⁷. The amount of transfected DNA per well was: mouse *Wnt1*, 5 ng; mouse *Fz8*, 1 ng; human *LRP6*, 3 ng; mouse *Wnt3a*, 1 ng; human *LRP6- Δ E(1-4)*, 5 ng; human β -catenin (*CTNNB1*), 10 ng; human *DVL1*, 5 ng; *Xenopus* *ck1 γ 1*, 1 ng; *Xenopus* *ck1 γ 1- Δ C*, 1 ng; *Xenopus* *ck1 γ 1^{K73R}*, 1 ng; *Xenopus* *ck1 γ 1^{D164N}*, 15 ng; human *LRP6* deletions and *LDLR- Δ N* mini-receptors, 10 ng. All transfections in HEK293T cells were performed with FuGene6 transfection reagent (Roche).

For *Drosophila* Wg reporter assays, SL2 cells were transfected in 12-well plates with 250 ng *LEF7-luc*, 80 ng *pRp128-Rluc* and 250 ng *pAc-lef-HA* (K. Bartscherer and M. Boutros, unpublished) together with 250 ng of either *pAc-wg* or *pMT-dsh* using Effectene (Qiagen). After 16 h, cells were added to dsRNA for gene silencing as described¹⁸. Dsh was induced with 500 μ M CuSO₄ 24 h before harvesting for the luciferase assay. Error bars show s.d. from mean of three independent samples. A *gish* PCR fragment was amplified from SL2 cell RNA using T7-linked primers (forward, 5'-GGCAGAACGTCAACAAACGTG-3'; reverse, 5'-GTTTGAATGTGAAGTGGCAGCG-3') and used for dsRNA production as described¹⁸. Primers used to amplify *arm* and *dsh* fragments are described online at <http://rnai.dkfz.de>.

BRET assays were performed as described¹⁹ in live HEK293T cells transfected in 24-well plates with 200 ng *LRP6-RL*, 10 ng *EYFP-ck1 γ 1*, 20 ng *EYFP-ck1 γ 1- Δ C* and 5 ng *EYFP*.

Embryos, *in situ* hybridization and RT-PCR. *In vitro* fertilization, embryo culture, staging, preparation of mRNA, microinjection and *in situ* hybridization were carried out as described²⁰.

PCR with reverse transcription (RT-PCR) was carried out as described²¹. PCR primers for *Xenopus laevis* *ck1 γ 1* were: forward, 5'-CCCTTGTGCCAGGCTCTATG-3'; reverse, 5'-CAAACTTGGGCCTAACAGCTCC-3'. Primers for *ck1 γ 2* were: forward, 5'-CTGAGGAAGTTGCCAGGAAGATGT-3'; reverse, 5'-GCCCAATAACTCAAGCACCATAG-3'. The sequence of the *ck1 γ 1* morpholino, targeting the translation start site of the *X. tropicalis* mRNA, is

5'-CTTTCCCTATTGTTATGGTCCATG-3'. It has one and two nucleotide mismatches with the two *X. laevis* *ck1γ1* (pseudo-)alleles.

Metabolic labelling and kinase assays. For *in vivo* labelling, HEK293T cells in 60-mm dishes were metabolically labelled 48 h after transfection with [³²P]orthophosphate (8 mCi per dish) for 6 h as described²². The membrane fraction from labelled cells was dissolved in 150 μl of lysis buffer (0.5% SDS, 10 mM DTT, 5 mM EDTA, 50 mM Tris-HCl, pH 8.5) (99 °C, 5 min). Proteins were alkylated in 50 mM iodoacetamide (37 °C, 40 min). Protein samples were diluted fivefold by the addition of 180 μl 20% NP40, 180 μl 5 M NaCl, 90 μl Tris-HCl, pH 7.5, and 950 μl H₂O. FLAG-LRP6 was immunoprecipitated using FLAG beads. After washing, half of the beads were eluted with Laemmli buffer. The remaining beads were subjected to λ-phosphatase treatment (30 °C, 30 min). Samples were separated by SDS-PAGE and analysed by autoradiography and immunoblot with the same membrane.

In vitro kinase assay with CK1γ and CK1γ-ΔC using peptide substrate (Fig. 2c) was performed as described²³. FLAG-tagged CK1γ1 or CK1γ1-ΔC was immunopurified from transfected HEK293T cells using FLAG beads and washed five times with lysis buffer containing 0.5 M NaCl. Prior to the kinase reaction, beads were washed two times with kinase buffer.

For *in vitro* kinase assay with LDLR-ΔN mini-receptors, constructs encoding FLAG-CK1γ1-ΔC and FLAG-tagged mini-receptors were transfected separately in HEK293T cells in 6-well plates. After 40 h, cells were lysed in 1% Triton buffer. Lysate from each mini-receptor was immunoprecipitated on FLAG beads overnight at 4 °C, with or without lysate from cells overexpressing FLAG-CK1γ1-ΔC. Beads were washed as before, and resuspended in 45 μl kinase buffer. Kinase reactions were carried out with 5 μl of [³²P]ATP (5 mCi mmol⁻¹) at 37 °C for 1 h. Reaction products were dissolved in Laemmli buffer (95 °C, 5 min) and supernatants analysed by autoradiography and immunoblot.

Antibodies, immunoprecipitation and immunoblots. Refer to the Supplementary Methods for details of these procedures.

Received 29 April; accepted 17 August 2005.

- Logan, C. Y. & Nusse, R. The Wnt Signaling Pathway in Development and Disease. *Annu. Rev. Cell Dev. Biol.* **20**, 781–810 (2004).
- Moon, R. T., Kohn, A. D., De Ferrari, G. V. & Kaykas, A. WNT and β-catenin signalling: diseases and therapies. *Nature Rev. Genet.* **5**, 691–701 (2004).
- He, X., Semenov, M., Tamai, K. & Zeng, X. LDL receptor-related proteins 5 and 6 in Wnt/β-catenin signalling: arrows point the way. *Development* **131**, 1663–1677 (2004).
- Belanger, C., Ansanay, H., Qanbar, R. & Bouvier, M. Primary sequence requirements for S-acylation of β₂-adrenergic receptor peptides. *FEBS Lett.* **499**, 59–64 (2001).
- Knippschild, U. *et al.* The casein kinase 1 family: participation in multiple cellular processes in eukaryotes. *Cell. Signal.* **17**, 675–689 (2005).
- Peters, J. M., McKay, R. M., McKay, J. P. & Graff, J. M. Casein kinase I transduces Wnt signals. *Nature* **401**, 345–350 (1999).
- Mao, B. *et al.* LDL-receptor-related protein 6 is a receptor for Dickkopf proteins. *Nature* **411**, 321–325 (2001).
- Hummel, T., Attix, S., Gunning, D. & Zipursky, S. L. Temporal control of glial cell migration in the *Drosophila* eye requires *gilgamesh*, *hedgehog*, and eye specification genes. *Neuron* **33**, 193–203 (2002).
- Xu, Y., Piston, D. W. & Johnson, C. H. A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. *Proc. Natl. Acad. Sci. USA* **96**, 151–156 (1999).
- Tamai, K. *et al.* A mechanism for Wnt coreceptor activation. *Mol. Cell* **13**, 149–156 (2004).
- Mao, J. *et al.* Low-density lipoprotein receptor-related protein-5 binds to Axin and regulates the canonical Wnt signalling pathway. *Mol. Cell* **7**, 801–809 (2001).
- Tolwinski, N. S. & Wieschaus, E. Rethinking WNT signalling. *Trends Genet.* **20**, 177–181 (2004).
- Liu, G., Bafico, A., Harris, V. K. & Aaronson, S. A. A novel mechanism for Wnt activation of canonical signalling through the LRP6 receptor. *Mol. Cell. Biol.* **23**, 5825–5835 (2003).
- Cong, F., Schweizer, L. & Varmus, H. Wnt signals across the plasma membrane to activate the β-catenin pathway by forming oligomers containing its receptors, Frizzled and LRP. *Development* **131**, 5103–5115 (2004).
- Carninci, P. *et al.* Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617–1630 (2000).
- Shibamoto, S. *et al.* Cytoskeletal reorganization by soluble Wnt-3a protein signalling. *Genes Cells* **3**, 659–670 (1998).
- Wu, W., Glinka, A., Delius, H. & Niehrs, C. Mutual antagonism between *dickkopf1* and *dickkopf2* regulates Wnt/β-catenin signalling. *Curr. Biol.* **10**, 1611–1614 (2000).
- Boutros, M. *et al.* Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303**, 832–835 (2004).
- Mao, B. *et al.* Kremen proteins are Dickkopf receptors that regulate Wnt/β-catenin signalling. *Nature* **417**, 664–667 (2002).
- Gawantka, V., Delius, H., Hirschfeld, K., Blumenstock, C. & Niehrs, C. Antagonizing the Spemann organizer: role of the homeobox gene *Xvent-1*. *EMBO J.* **14**, 6268–6279 (1995).
- Dosch, R., Gawantka, V., Delius, H., Blumenstock, C. & Niehrs, C. Bmp-4 acts as a morphogen in dorsoventral mesoderm patterning in *Xenopus*. *Development* **124**, 2325–2334 (1997).
- Reimer, T. *et al.* Phosphorylation of the N-terminal domain regulates subcellular localization and DNA binding properties of the peptidyl-prolyl *cis/trans* isomerase hPar14. *J. Mol. Biol.* **330**, 955–966 (2003).
- Zhai, L. *et al.* Casein kinase 1γ subfamily. Molecular cloning, expression, and characterization of three mammalian isoforms and complementation of defects in the *Saccharomyces cerevisiae* YCK genes. *J. Biol. Chem.* **270**, 12717–12724 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Bartscherer and M. Boutros for providing reagents and help with the *Drosophila* cell experiments, H. Clevers, R. Grosschedl, X. He, R. Moon, J. Nathans and R. Nusse for reagents, and H. Spring for confocal microscopy. We also thank S. Cohen and M. Boutros for comments on the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft.

Author Information The sequence of *Xenopus* *ck1γ2* has been deposited into GenBank (accession number DQ185136). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.N. (Niehrs@dkfz-heidelberg.de) or G.D. (g.davidson@dkfz-heidelberg.de).

A dual-kinase mechanism for Wnt co-receptor phosphorylation and activation

Xin Zeng¹, Keiko Tamai¹, Brad Doble³, Shitao Li¹, He Huang¹, Raymond Habas^{1†}, Heidi Okamura^{2†}, Jim Woodgett³ & Xi He¹

Signalling by the Wnt family of secreted lipoproteins has essential functions in development and disease¹. The canonical Wnt/ β -catenin pathway requires a single-span transmembrane receptor, low-density lipoprotein (LDL)-receptor-related protein 6 (LRP6)^{2–4}, whose phosphorylation at multiple PPPSP motifs is induced upon stimulation by Wnt and is critical for signal transduction⁵. The kinase responsible for LRP6 phosphorylation has not been identified. Here we provide biochemical and genetic evidence for a ‘dual-kinase’ mechanism for LRP6 phosphorylation and activation. Glycogen synthase kinase 3 (GSK3), which is known for its inhibitory role in Wnt signalling through the promotion of β -catenin phosphorylation and degradation, mediates the phosphorylation and activation of LRP6. We show that Wnt induces sequential phosphorylation of LRP6 by GSK3 and casein kinase 1, and this dual phosphorylation promotes the engagement of LRP6 with the scaffolding protein Axin. We show further that a membrane-associated form of GSK3, in contrast with cytosolic GSK3, stimulates Wnt signalling and *Xenopus* axis duplication. Our results identify two key kinases mediating Wnt co-receptor activation, reveal an unexpected and intricate logic of Wnt/ β -catenin signalling, and illustrate GSK3 as a genuine switch that dictates both on and off states of a pivotal regulatory pathway.

Canonical Wnt signalling operates through regulating the phosphorylation and degradation of the transcription co-activator β -catenin^{1,6}. Without stimulation by Wnt, β -catenin is assembled into the Axin complex, in which β -catenin is sequentially phosphorylated by casein kinase 1 (CK1) and GSK3 and earmarked for degradation^{7–9}. Stimulation by Wnt leads to the inhibition of β -catenin phosphorylation and degradation. This signal transduction is initiated at the plasma membrane by two distinct receptors, a Frizzled serpentine receptor and LRP6 or LRP5, which together may form a Wnt-induced Frizzled–LRP6 (or LRP5) complex^{3,6,10–12}. Although the mechanism by which this receptor pair initiates signalling remains to be understood, Wnt-induced LRP6 phosphorylation at a PPPSP motif, which is reiterated five times in the cytoplasmic domain of LRP5/6 (Fig. 1a), has a crucial function⁵. (For simplicity we use PPPSP to represent PPPSP or PPPTP). Indeed, LRP6 mutants lacking these motifs or harbouring substitutions at the S/T residues are inactive in signalling^{5,13}. Conversely, a single PPPSP motif, on transfer to a heterologous receptor, is sufficient to initiate β -catenin signalling⁵. Because the phosphorylated PPPSP motif mediates LRP5/6–Axin interaction^{5,14}, we proposed a model⁵ in which Wnt-induced LRP6 phosphorylation promotes the recruitment of Axin by LRP6, thereby permitting the Wnt receptor complex to regulate β -catenin phosphorylation directly¹⁴.

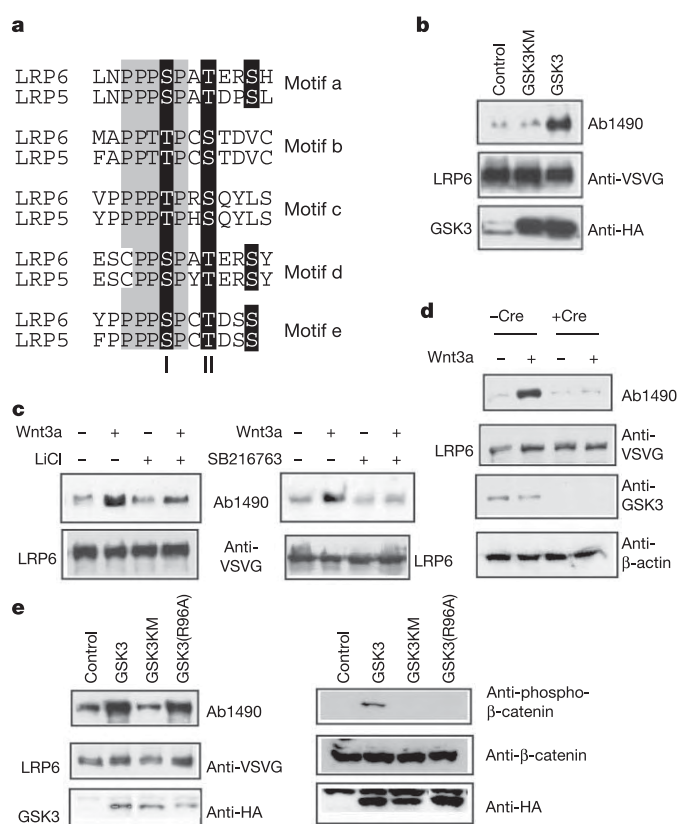


Figure 1 | GSK3 involvement in PPPSP phosphorylation. **a**, Five PPPSP motifs (motifs a–e) in LRP5/6. Site I, site II and another potential phosphorylation site are highlighted. **b**, GSK3 β , but not the kinase-dead GSK3 β KM, promoted LRP6 phosphorylation. LRP6 PPPSP phosphorylation *in vivo* was examined in 293T cells or MEFs stably expressing LRP6 (VSVG-tagged)⁵ in this and other figures. **c**, Wnt3a-induced LRP6 PPPSP phosphorylation in MEFs was inhibited by LiCl (50 mM) or SB216763 (3 or 30 μ M). **d**, Wnt3a-induced LRP6 PPPSP phosphorylation was abolished in MEFs lacking *Gsk3 α* and *Gsk3 β* . *Gsk3 α* alleles were deleted by means of Cre expression in *Gsk3 β ^{-/-};**Gsk3 α ^{flox/flox}* MEFs. β -Actin was used as a loading control. **e**, GSK3 β (R96A) phosphorylated LRP6 PPPSP motif, but not β -catenin S33/S37/T41, detected with a phosphorylation-specific antibody⁷. GSK3 β KM phosphorylated neither LRP6 nor β -catenin.

¹Neurobiology Program, Children's Hospital Boston, Department of Neurology, and ²CBR Institute and Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. ³Ontario Cancer Institute/Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario, Canada M5G 2M9. [†]Present addresses: AVEO Pharmaceuticals, Inc., 75 Sidney Street, Cambridge, Massachusetts 02139, USA (H.O.); Department of Biochemistry, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson School of Medicine, Piscataway, New Jersey 08854, USA (R.H.).

We tested whether the PPPSP kinase might associate with LRP6 and/or Axin, because Axin overexpression enhanced LRP6 PPPSP phosphorylation⁵. Glutathione *S*-transferase (GST)–LRP6C, a fusion protein between GST and the LRP6 cytoplasmic domain, and Axin each precipitated from 293T cell extracts an associated PPPSP kinase (Supplementary Figs 1 and 2). We suspected the involvement of GSK3. First, we isolated GSK3 α in a yeast two-hybrid screen, using LRP6C as bait, and confirmed this interaction in yeast (not shown), indicating that GSK3 might bind LRP6. Second, GSK3 associates with Axin. Third, GSK3 is a proline-directed kinase that prefers (S/T)P for phosphorylation and was predicted by Scansite 2.0 (ref. 15) to phosphorylate the PPPSP motif. Finally, GSK3 overexpression promotes LRP5–Axin interaction¹⁴, which is consistent with the possibility that GSK3 might do so through the phosphorylation of LRP5/6. Indeed, the PPPSP kinase associated with GST–LRP6C or Axin was inhibited *in vitro* by LiCl or SB216763, two GSK3 inhibitors^{16,17} (Supplementary Figs 1 and 2). *In vivo*, GSK3 overexpression promoted LRP6 PPPSP phosphorylation (Fig. 1b),

whereas LiCl and SB216763 prevented Wnt-induced LRP6 PPPSP phosphorylation (Fig. 1c). Wnt-induced LRP6 phosphorylation was abolished in mouse embryo fibroblasts (MEFs) harbouring genetic deletions of the *Gsk3 α* and *Gsk3 β* genes (Fig. 1d). These data show that GSK3 is involved in LRP6 PPPSP phosphorylation.

Phosphorylation of many substrates by GSK3 requires a priming kinase¹⁷; for β -catenin this is CK1 (refs 7–9). None of the S/T residues in LRP6 PPPSP motifs and flanking regions (Fig. 1a) conforms to the priming phosphorylation consensus^{7,17}, indicating a possibly priming-independent GSK3 phosphorylation. We tested this possibility with GSK3 β (R96A), which has alanine substituted for arginine 96 and is incapable of recognizing priming phosphorylation^{18–21}. GSK3 β (R96A) was inactive in β -catenin phosphorylation²¹ but active in LRP6 PPPSP phosphorylation (Fig. 1e). GSK3 therefore phosphorylates LRP6 and β -catenin by means of distinct mechanisms.

A single PPPSP motif on transfer to a truncated LDL receptor (LDLR Δ N) is sufficient to activate the Wnt pathway⁵. During characterization of this motif, we noticed a second S/T (site II)

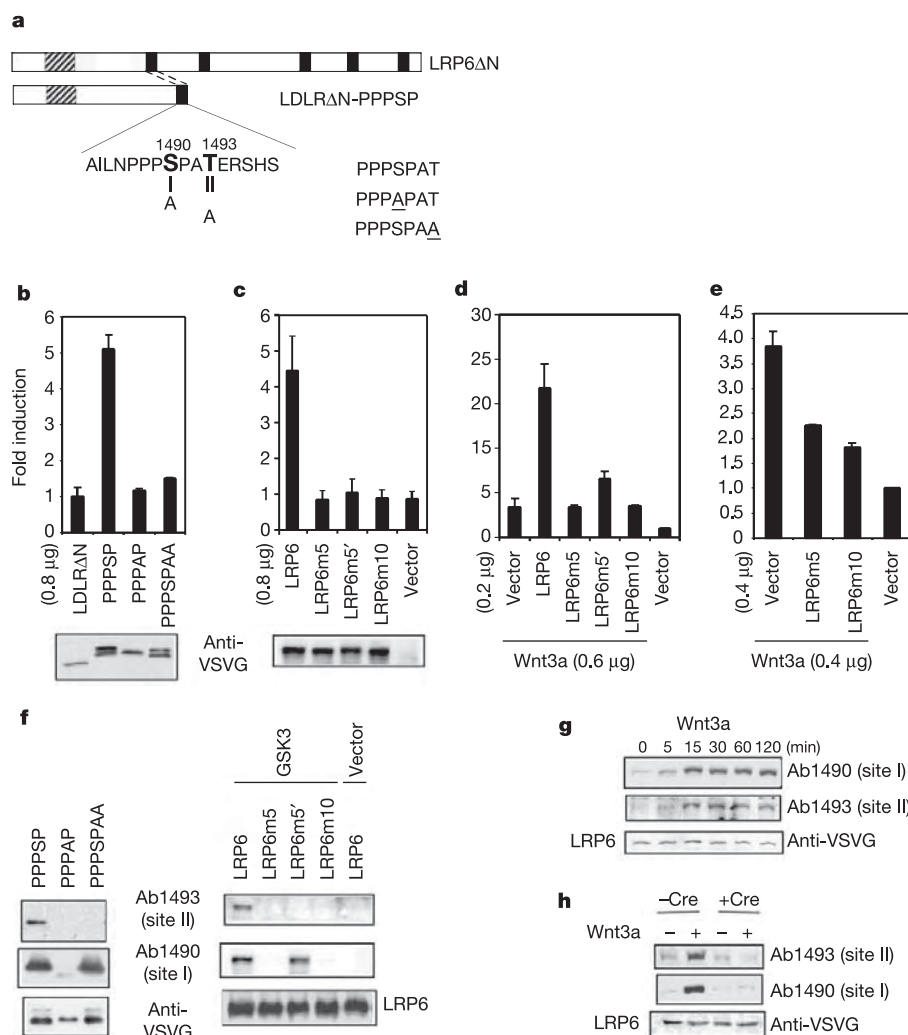


Figure 2 | Sequential phosphorylation of the PPPSPxS motif is induced by Wnt3a and required for LRP6 signalling. **a**, Scheme of five PPPSPxS motifs (black rectangles) in LRP6 Δ N and of LDLR Δ N-PPPSP. Site I, site II and their alanine substitutions are illustrated. Hatching indicates the transmembrane domain. **b–e**, TCF/ β -catenin reporter assays. **b**, **c**, Alanine substitution at site II or site I diminished the activity of LDLR Δ N-PPPSP (**b**) and LRP6 (**c**). Protein expression level was examined by means of the VSVG tag. The site II mutation did not affect the mobility shift⁵, indicating normal site I phosphorylation (**b**). **d**, **e**, LRP6m5 and LRP6m10 did not synergize with Wnt3a (**d**) but inhibited signalling by Wnt3a (**e**) or by Wnt8 in *Xenopus*

embryos⁵ (not shown). LRP6m5' had diminished ability to synergize with Wnt3a (**d**) but did not inhibit Wnt3a signalling. **f**, Sequential phosphorylation of site I and site II in LDLR Δ N-PPPSP and LRP6 (see Supplementary Fig. 4). Left, LDLR Δ N-PPPSP/derivatives. Ab1490 and Ab1493 specifically recognized the phosphorylated, more slowly migrating, band. Right, GSK3 β promoted LRP6 phosphorylation at site I and site II. **g**, **h**, LRP6 phosphorylation at site I and site II was induced by Wnt3a CM in MEFs (**g**) but not in MEFs lacking *Gsk3 α* and *Gsk3 β* (**h**). Luciferase activity is presented as fold induction compared with the control (mean $\pm$ s.d.).

three amino-acid residues downstream of the PPPSP phosphorylation site (site I, Fig. 2a). Site II is present in all PPPSP motifs (Fig. 1a) and is conserved in *Drosophila* Arrow⁴, the orthologue of LRP5/6; it was included in the flanking residues in the LDLRΔN-PPPSP construct⁵ (Fig. 2a). Replacing the site II threonine with alanine in LDLRΔN-PPPSP, like replacing the site I serine with alanine, diminished its signalling activity in a T-cell factor (TCF)/β-catenin reporter assay (Fig. 2b). We demonstrated the importance of site II S/T residues in the full-length LRP6. LRP6m5', which harbours an alanine substitution of all five S/T residues at site II, was much less active than LRP6, alone or in synergy with Wnt3a, in the TCF/β-catenin reporter assay (Fig. 2c, d). LRP6m5 and LRP6m10, which have an alanine substitution of five site-I S/T residues⁵ and of five site-I plus five site-II S/T residues, respectively, were inactive alone or in synergy with Wnt3a (Fig. 2c, d). When expressed more highly they each behaved as a dominant-negative mutant and suppressed Wnt3a signalling (Fig. 2e). Therefore site I is essential for, and site II augments, Wnt/LRP6 signalling.

Site II, like site I, was phosphorylated (Fig. 2f, g), as detected by means of a phospho-specific antibody (Ab1493, for phosphorylated T1493 in LRP6; Supplementary Fig. 3). LDLRΔN-PPPSP, which is constitutively active, showed constitutive phosphorylation at site I and site II (Fig. 2f). Wnt3a induced rapid LRP6 phosphorylation at site I and site II (Fig. 2g). Phosphorylation of the extended PPPSPxS motif therefore correlates fully with Wnt/LRP6 signalling.

In both LDLRΔN-PPPSP and LRP6, mutation at site I abolished phosphorylation at both site I and site II, whereas mutation at site II had no effect on site I phosphorylation (Fig. 2f). The simplest

explanation is that the phosphorylation of site I by GSK3 both primes and is required for site II phosphorylation. Indeed, GSK3 overexpression increased LRP6 phosphorylation at site I and site II (Fig. 2f), whereas a lack of GSK3 prevented Wnt-induced LRP6 phosphorylation at either site I or site II (Fig. 2h).

We considered that the site II kinase is not GSK3 because the order and spacing between site II and site I phosphorylation are not what GSK3 requires. In fact site II resembles the CK1 phosphorylation consensus (S/T)*xx(S/T), in which (S/T)* is a priming phosphorylated residue and xx is any two-amino acid combination²². Seven mammalian CK1 members, namely α, α2, δ, ε, γ1, γ2 and γ3, exist. CK1α, CK1δ and CK1ε have been implicated in Wnt/β-catenin signalling with apparently opposing roles: CK1α is a priming kinase for β-catenin phosphorylation/degradation⁷, whereas CK1δ and CK1ε activate β-catenin signalling^{23,24} (also see ref. 8) by mechanisms that have yet to be fully understood. CK1γ3 overexpression can also stimulate, modestly, Wnt/β-catenin signalling²⁵. Because of the multiplicity of the CK1 family, we used the specific dominant-negative (DN) CK1 mutants DN-CK1α and DN-CK1δ. DN-CK1α inhibits CK1α but not CK1δ/ε, whereas DN-CK1δ inhibits CK1δ/ε but not CK1α (S.L. and X.H., unpublished observations). DN-CK1α plus DN-CK1δ, but neither alone, prevented Wnt-induced phosphorylation at site II but not at site I (Fig. 3a). Thus CK1 probably phosphorylates site II after phosphorylation of the priming site I by GSK3, and CK1α (and CK1α2) and CK1δ/ε may have redundant roles in the process.

We showed that Axin preferentially binds to a phosphorylated PPPSP motif in LDLRΔN-PPPSP and LRP6 (ref. 5), which are also phosphorylated at site II (Fig. 2f, g). The site II mutation diminished Axin binding to the PPPSPxS motif, whereas the site I mutation abolished it (Fig. 3b). Phosphorylation at both site I and site II is therefore required for Axin recognition *in vivo*. We reconstituted GST-LRP6C phosphorylation *in vitro* by means of recombinant GSK3 and/or CK1 (Fig. 3c). GSK3 and CK1 phosphorylated site I and site II, respectively, and GSK3 phosphorylation of site I enhanced CK1 phosphorylation of site II (Fig. 3c). GST-LRP6C phosphorylated by GSK3 plus CK1, but not by GSK3 or CK1 alone, exhibited strong Axin-binding capability (Fig. 3c). This binding of Axin required the intact PPPSPxS motifs, because GST-LRP6Cm5 (site I mutated in all five motifs), GST-LRP6Cm5' (site II mutated in all five motifs) and GST-LRP6Cm10 (site I plus site II mutated in all five motifs) were defective in Axin interaction (Fig. 3d). Phosphorylation of the PPPSPxS motif by GSK3 and CK1 is therefore essential for the recruitment of Axin by LRP6.

The phosphorylation of LRP6 by CK1 offers a mechanism to account, in part, for a documented activating role of CK1 in Wnt/β-catenin signalling. However, our results that GSK3 is required for LRP6 phosphorylation and activation depart drastically from the dogma that GSK3 is solely an inhibitor of Wnt signalling by means of its role in β-catenin phosphorylation and degradation. We proposed that different pools of GSK3 have distinct functions in the Wnt pathway. In this model, cytosolic GSK3 phosphorylates β-catenin together with CK1 and thus antagonizes Wnt/β-catenin signalling, whereas plasma-membrane-associated GSK3 phosphorylates LRP6 together with CK1 in response to Wnt stimulation and activates Wnt/β-catenin signalling. We found significant amounts of membrane-associated GSK3 and CK1, which did not change appreciably with or without Wnt stimulation (Fig. 4a). We then generated mGSK3, a GSK3 mutant that targets GSK3 to the plasma membrane by means of a mini-transmembrane unit, LDLRΔN⁵. mGSK3, like GSK3, stimulated LRP6 phosphorylation (Fig. 4b). However, in sharp contrast to GSK3, which antagonized LRP6 signalling, mGSK3 synergized with LRP6 in activating the TCF/β-catenin reporter (Fig. 4c). mGSK3 did not synergize with LRP6m5 (Fig. 4d) and therefore depended on the PPPSP motif for its stimulatory activity. In *Xenopus* embryos, mGSK3 plus LRP6 induced axis duplication and the expression of Xnr3 (Fig. 4e, f), a

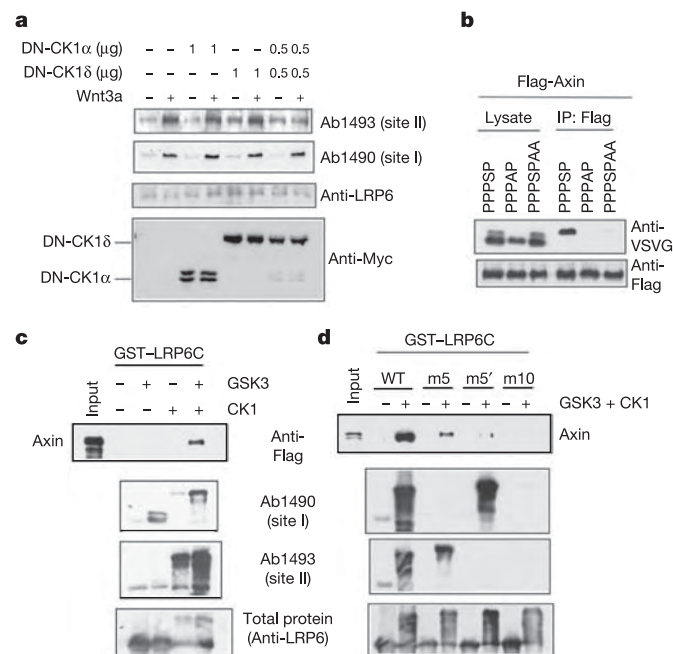


Figure 3 | Site I and site II phosphorylation by GSK3 and CK1 promotes LRP6 recruitment of Axin. **a**, DN-CK1α plus DN-CK1δ, but neither alone, prevented Wnt3a-induced phosphorylation of site II but not that of site I. *Cklε*^{-/-} MEFs (Supplementary Fig. 5) were used. **b**, Alanine substitution at site I or site II in the PPPSPxS motif prevented the binding of Axin. Axin was co-expressed with LDLRΔN-PPPSP or mutants. Axin co-precipitated the phosphorylated motif but precipitated neither the unphosphorylated one (fast migrating) nor the site I or site II mutant, which had normal site I phosphorylation. **c**, GST-LRP6C binding to Axin *in vitro* required phosphorylation by recombinant GSK3β plus CK1ε. Phosphorylated GST-LRP6C was examined for phosphorylation and used to precipitate cell extracts expressing Flag-tagged Axin. **d**, Axin binding to phosphorylated GST-LRP6C required site I and site II. Axin showed much diminished binding to GST-LRP6Cm5, GST-LRP6Cm5' and GST-LRP6Cm10. WT, wild type.

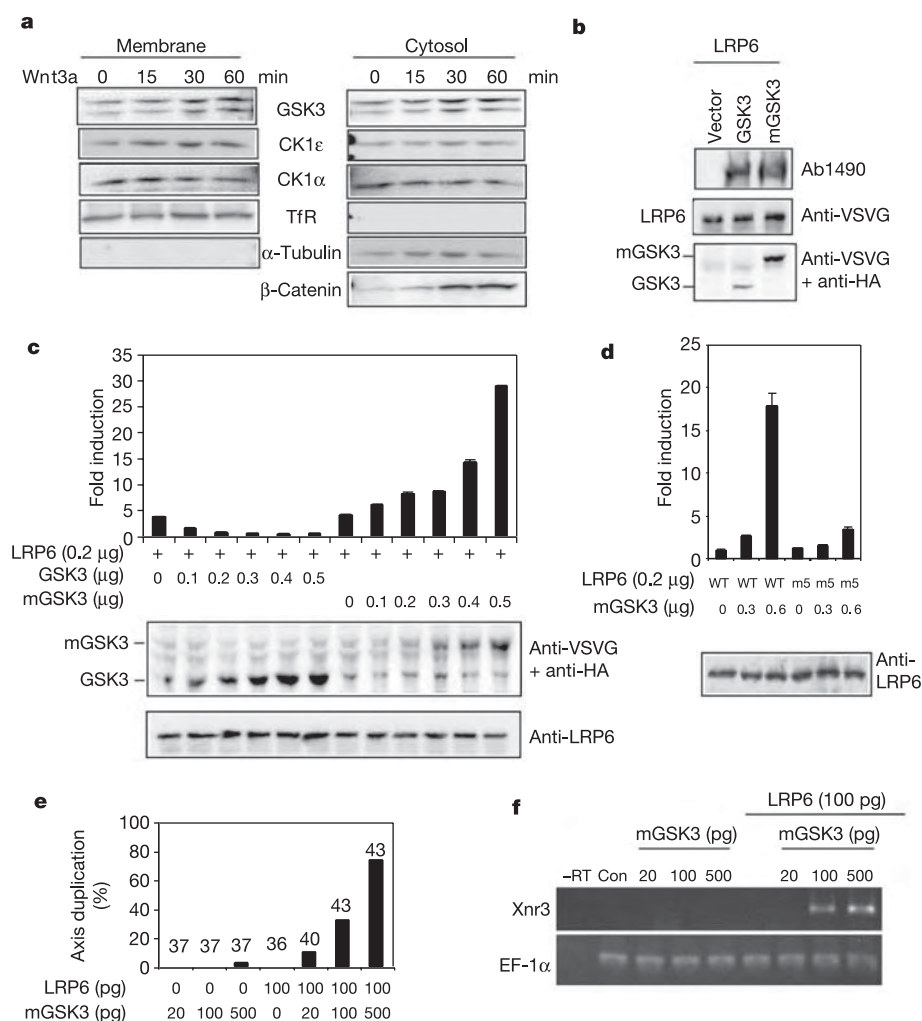


Figure 4 | Membrane-associated GSK3 phosphorylates LRP6 and activates LRP6 and TCF/β-catenin signalling. **a**, Membrane-associated GSK3 α/β and CK1 α/ϵ is independent of Wnt3a stimulation. GSK3 α/β (10%) and CK1 α/ϵ (20%) were detected in the membrane fraction, which was marked by the transferrin receptor (TfR) and a lack of α -tubulin. Wnt3a stabilized cytosolic β -catenin. **b**, mGSK3 β , like GSK3 β , promoted LRP6 PPPSP phosphorylation. mGSK3 β (VSVG-tagged) and GSK3 β (HA-tagged) were detected with the two antibodies. **c**, **d**, TCF/β-catenin reporter assays. mGSK3 β activated, whereas GSK3 β inhibited, LRP6 signalling (**c**). mGSK3 β

did not synergize with LRP6m5; the slight stimulation seen might be due to the endogenous LRP5/6 (**d**). WT, wild type. **e**, **f**, mGSK3 β synergized with LRP6 to induce axis duplication (**e**) and Xnr3 expression (**f**) in *Xenopus* embryos. RNA doses injected are per embryo. The number above each bar in **e** indicates the number of embryos injected. **f**, Reverse transcriptase-mediated polymerase chain reaction assay. EF-1 α , loading control; –RT, without reverse transcriptase; con, uninjected embryos. Luciferase activity is presented as fold induction compared with the control (mean $\pm$ s.d.).

TCF/β-catenin target. GSK3 in the proximity of the plasma membrane therefore mediates Wnt/LRP6 signalling by phosphorylating the PPPSP motif.

We have provided biochemical, genetic and functional evidence that GSK3 and CK1 sequentially phosphorylate PPPSPxS motifs in LRP6 on stimulation by Wnt, and this dual-kinase phosphorylation underlies LRP6 activation by promoting Axin recruitment through phosphorylated LRP6. The critical function of GSK3 in mediating Wnt/β-catenin signalling is probably performed by membrane-associated GSK3, which acts to antagonize β-catenin phosphorylation and degradation by cytosolic GSK3. These opposing functions of GSK3 in a single pathway are unprecedented but may be separated spatially and temporally. How Wnt induces LRP6 phosphorylation by GSK3 and CK1 remains unknown. Because GSK3 and CK1 exhibit membrane association regardless of Wnt stimulation, Wnt may control the access of GSK3 to LRP6. Alternatively, Wnt may inhibit a phosphatase that counteracts the constitutive LRP6 phosphorylation by GSK3 and CK1. Our study further implies that in *Drosophila*, Arrow is probably activated by GSK3/Zeste-White 3/Shaggy or another GSK3-related protein (see refs 11, 26, 27).

Sequential phosphorylation of LRP6 and β-catenin mirror each other; that is, they are regulated by Wnt in opposite directions and use GSK3 and CK1 as the mutual priming kinase. However, in both cases GSK3 phosphorylation seems to be the main regulatory step, whereas CK1 phosphorylation may be constitutive. GSK3 uses distinct mechanisms to phosphorylate β-catenin (a primed substrate) and LRP6 (a non-primed substrate). Our findings also uncover further striking parallels between the Wnt and Hedgehog (Hh) pathways. In the absence of Hh, protein kinase A and CK1 (and GSK3) sequentially phosphorylate transcription factor Ci for processing Ci into a repressor, thus suppressing Hh signalling. On stimulation of Hh, protein kinase A and CK1 sequentially phosphorylate the cytoplasmic domain of the Smoothened receptor, leading to the inhibition of Ci phosphorylation and thus to the activation of Hh signalling^{28–30}.

METHODS

Plasmids. VSVG (YTDIEMNRLGK)-tagged LRP6, LRP6m5, LDLRΔN-PPPSP and LDLRΔN-PPPAP, Flag-tagged Axin, haemagglutinin (HA)-tagged GSK3 β and GSK3 β KM in pCS2+ have been described³⁵. VSVG-tagged LRP6m5' and

LRP6m10 were generated by sequential mutagenesis from VSVG-LRP6 in pCS2+ with the use of the QuickChange Kit (Stratagene). HA-GSK3 β (R96A) was generated by site-directed mutagenesis of HA-GSK3 β in pCS2+. mGSK3 was generated by replacing the last 11 amino acid residues of VSVG-tagged LDLRAN⁵ with the coding sequence of Gsk3 β . Myc-tagged DN-CKI α and DN-CKI δ will be described elsewhere (details are available from the authors). GST-LRP6C and GST-LRP6N1 are in pGEX with the GST coding region fused in-frame to the entire (residues 1397–1613) and a partial (residues 1397–1534) cytoplasmic domain of LRP6, respectively. Details of the plasmids are available from the authors on request.

Antibodies. The polyclonal antibody Ab1490 has been described⁵. Polyclonal Ab1493 was generated similarly except that the synthetic phosphopeptide CLNPPSPAT*ERSH (in which T* represents phosphothreonine) was used as the immunogen. Other antibodies were used in accordance with the manufacturers' instructions: anti- β -catenin, C2206 (Sigma); anti-phospho- β -catenin, S33/S37/T41 (Cell Signaling); anti-VSVG, V4888 (Sigma); anti- α -tubulin, sc8053 (Santa Cruz); anti-HA, Sc7392 (Santa Cruz); anti-Myc, A14 (Santa Cruz); anti-Flag, M2 (Sigma); anti-GSK3, 4G-1E (Upstate); anti-transferrin receptor, 13-6800 (Zymed); anti-CKI ϵ , 610445 (BD Transduction Lab); and anti-CKI α , sc-6477 (Santa Cruz).

Cell culture, Wnt induction, transfections and reporter assays. 293T cells were used unless MEFs were specifically mentioned. 293T cells or MEFs stably expressing VSVG-LRP6 or Flag-tagged Axin were established by means of puromycin selection after co-transfection of LRP6 or Axin with the pBABE puro plasmid. These lines have essentially the same Wnt response as parental cells (not shown). MEFs and 293T cells were maintained in DMEM medium supplemented with 10% FBS and 1% penicillin-streptomycin-glutamine. Transfections were conducted with Fugene6 (Roche). Luciferase assays were performed in 293T cells in duplicates as described previously⁵, with 0.3 μ g of TCF-luciferase reporter plasmid plus 50 ng of *Renilla* luciferase plasmid (internal control) per well in the 12-well plate. Each experiment was repeated at least three times. Luciferase activity, presented as fold induction compared with the control, is shown as mean $\pm$ s.d. For induction of LRP6 phosphorylation, cells were treated for 1 h with Wnt3a CM (collected from mouse L cells stably expressing Wnt3a, from ATCC) unless indicated otherwise.

Cre-mediated deletion of Gsk3 α alleles in Gsk3 $\beta^{-/-}$;Gsk3 $\alpha^{lox/lox}$ MEFs. Gsk3 $\alpha^{-/-}$;Gsk3 $\beta^{-/-}$ MEFs failed to survive after several passages. Gsk3 $\beta^{-/-}$;Gsk3 $\alpha^{lox/lox}$ MEFs (Supplementary Information) were freshly infected with an adenoviral Cre expression system to delete both alleles of the Gsk3 α gene and were maintained for a further 3 days for maximal turnover of the residual GSK3 α protein before each experiment. After Cre expression, these MEFs exhibit normal response to epidermal growth factor and contain high levels of β -catenin protein as expected (not shown). Control MEFs infected with the Cre expression system exhibited normal Wnt3a induction of LRP6 phosphorylation (not shown).

Manipulations of *Xenopus laevis* embryos. Synthetic RNAs were transcribed *in vitro* with MessageMachine (Ambion) and were injected into the animal region at the two-cell stage for Xnr3 induction in animal pole explants or ventrally at the four-cell stage for axis duplication. Procedures for embryo staging, injection and animal pole explants have been described³.

Received 30 June; accepted 1 September 2005.

- Logan, C. Y. & Nusse, R. The Wnt signalling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.* **20**, 781–810 (2004).
- Pinson, K. I., Brennan, J., Monkley, S., Avery, B. J. & Skarnes, W. C. An LDL-receptor-related protein mediates Wnt signalling in mice. *Nature* **407**, 535–538 (2000).
- Tamai, K. *et al.* LDL-receptor-related proteins in Wnt signal transduction. *Nature* **407**, 530–535 (2000).
- Wehrli, M. *et al.* arrow encodes an LDL-receptor-related protein essential for Wingless signalling. *Nature* **407**, 527–530 (2000).
- Tamai, K. *et al.* A mechanism for Wnt coreceptor activation. *Mol. Cell* **13**, 149–156 (2004).
- He, X., Semenov, M., Tamai, K. & Zeng, X. LDL receptor-related proteins 5 and 6 in Wnt/ β -catenin signalling: arrows point the way. *Development* **131**, 1663–1677 (2004).
- Liu, C. *et al.* Control of β -catenin phosphorylation/degradation by a dual-kinase mechanism. *Cell* **108**, 837–847 (2002).
- Amit, S. *et al.* Axin-mediated CKI phosphorylation of β -catenin at Ser 45: a molecular switch for the Wnt pathway. *Genes Dev.* **16**, 1066–1076 (2002).

- Yanagawa, S. *et al.* Casein kinase I phosphorylates the Armadillo protein and induces its degradation in *Drosophila*. *EMBO J.* **21**, 1733–1742 (2002).
- Semenov, M. V. *et al.* Head inducer Dickkopf-1 is a ligand for Wnt coreceptor LRP6. *Curr. Biol.* **11**, 951–961 (2001).
- Tolwinski, N. S. *et al.* Wg/Wnt signal can be transmitted through arrow/LRP5,6 and Axin independently of Zw3/Gsk3 β activity. *Dev. Cell* **4**, 407–418 (2003).
- Cong, F., Schweizer, L. & Varmus, H. Wnt signals across the plasma membrane to activate the β -catenin pathway by forming oligomers containing its receptors, Frizzled and LRP. *Development* **131**, 5103–5115 (2004).
- Brennan, K., Gonzalez-Sancho, J. M., Castelo-Soccio, L. A., Howe, L. R. & Brown, A. M. Truncated mutants of the putative Wnt receptor LRP6/Arrow can stabilize β -catenin independently of Frizzled proteins. *Oncogene* **23**, 4873–4884 (2004).
- Mao, J. *et al.* Low-density lipoprotein receptor-related protein-5 binds to Axin and regulates the canonical Wnt signalling pathway. *Mol. Cell* **7**, 801–809 (2001).
- Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. Scansite 2.0: Proteome-wide prediction of cell signalling interactions using short sequence motifs. *Nucleic Acids Res.* **31**, 3635–3641 (2003).
- Coghlan, M. P. *et al.* Selective small molecule inhibitors of glycogen synthase kinase-3 modulate glycogen metabolism and gene transcription. *Chem. Biol.* **7**, 793–803 (2000).
- Doble, B. W. & Woodgett, J. R. GSK-3: tricks of the trade for a multi-tasking kinase. *J. Cell Sci.* **116**, 1175–1186 (2003).
- Frame, S., Cohen, P. & Biondi, R. M. A common phosphate binding site explains the unique substrate specificity of GSK3 and its inactivation by phosphorylation. *Mol. Cell* **7**, 1321–1327 (2001).
- Dajani, R. *et al.* Crystal structure of glycogen synthase kinase 3 β : structural basis for phosphate-primed substrate specificity and autoinhibition. *Cell* **105**, 721–732 (2001).
- ter Haar, E. *et al.* Structure of GSK3 β reveals a primed phosphorylation mechanism. *Nature Struct. Biol.* **8**, 593–596 (2001).
- Hagen, T., Di Daniel, E., Culbert, A. A. & Reith, A. D. Expression and characterization of GSK-3 mutants and their effect on β -catenin phosphorylation in intact cells. *J. Biol. Chem.* **277**, 23330–23335 (2002).
- Marin, O. *et al.* A noncanonical sequence phosphorylated by casein kinase 1 in β -catenin may play a role in casein kinase 1 targeting of important signalling proteins. *Proc. Natl Acad. Sci. USA* **100**, 10193–10200 (2003).
- Peters, J. M., McKay, R. M., McKay, J. P. & Graff, J. M. Casein kinase I transduces Wnt signals. *Nature* **401**, 345–350 (1999).
- Sakanaka, C., Leong, P., Xu, L., Harrison, S. D. & Williams, L. T. Casein kinase Ie in the Wnt pathway: regulation of β -catenin function. *Proc. Natl Acad. Sci. USA* **96**, 12548–12552 (1999).
- McKay, R. M., Peters, J. M. & Graff, J. M. The casein kinase I family in Wnt signalling. *Dev. Biol.* **235**, 388–396 (2001).
- Cliffe, A., Hamada, F. & Bienz, M. A role of Dishevelled in relocating Axin to the plasma membrane during Wingless signalling. *Curr. Biol.* **13**, 960–966 (2003).
- Ruel, L., Pantescio, V., Lutz, Y., Simpson, P. & Bourouis, M. Functional significance of a family of protein kinases encoded at the shaggy locus in *Drosophila*. *EMBO J.* **12**, 1657–1669 (1993).
- Jia, J., Tong, C., Wang, B., Luo, L. & Jiang, J. Hedgehog signalling activity of Smoothened requires phosphorylation by protein kinase A and casein kinase I. *Nature* **432**, 1045–1050 (2004).
- Zhang, C., Williams, E. H., Guo, Y., Lum, L. & Beachy, P. A. Extensive phosphorylation of Smoothened in Hedgehog pathway activation. *Proc. Natl Acad. Sci. USA* **101**, 17900–17907 (2004).
- Apionishev, S., Katanayeva, N. M., Marks, S. A., Kalderon, D. & Tomlinson, A. *Drosophila* Smoothened phosphorylation sites essential for Hedgehog signal transduction. *Nature Cell Biol.* **7**, 86–92 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Semenov and B. MacDonald for suggestions and discussion. This work is supported in part by a grant from the NIH to X.H., who is a W. M. Keck Foundation Distinguished Young Scholar and a Leukemia and Lymphoma Society Scholar. X.Z., K.T., H.H. and R.H. are or were in part supported by postdoctoral fellowships from the Children's Hospital Boston, Uehara Memorial Foundation (Japan) and CIHR (Canada), and a training grant from the NIH, respectively. B.D. and J.W. are supported by the CIHR. J.W. is a Howard Hughes Medical Institute International Scholar. H.O. was in part supported by a grant from the NIH to Anjana Rao.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to X.H. (xi.he@childrens.harvard.edu).

LETTERS

The importance of sequence diversity in the aggregation and evolution of proteins

Caroline F. Wright¹, Sarah A. Teichmann², Jane Clarke³ & Christopher M. Dobson¹

Incorrect folding of proteins, leading to aggregation and amyloid formation, is associated with a group of highly debilitating medical conditions^{1,2} including Alzheimer's disease and late-onset diabetes. The issue of how unwanted protein association is normally avoided in a living system is particularly significant in the context of the evolution of multidomain proteins, which account for over 70% of all eukaryotic proteins³, where the effective local protein concentration in the vicinity of each domain is very high. Here we describe the aggregation kinetics of multidomain protein constructs of immunoglobulin domains and the ability of different homologous domains to aggregate together. We show that aggregation of these proteins is a specific process and that the efficiency of coaggregation between different domains decreases markedly with decreasing sequence identity. Thus, whereas immunoglobulin domains with more than about 70% identity are highly prone to coaggregation, those with less than 30–40% sequence identity do not detectably interact. A bioinformatics analysis of consecutive homologous domains in large multidomain proteins shows that such domains almost exclusively have sequence identities of less than 40%, in other words below the level at which coaggregation is likely to be efficient. We propose that such low sequence identities could have a crucial and general role in safeguarding proteins against misfolding and aggregation.

Protein aggregation is a widespread problem both *in vitro*, in biotechnology and biomedical research, and *in vivo*, in pathological conditions. Indeed, both systemic and organ-specific diseases are associated with the deposition in tissue of highly structured protein aggregates known as amyloid fibrils or plaques^{1,2}. The generic and largely sequence-independent character of such structures might suggest that protein aggregation is a relatively non-specific coagulation of partially or completely unfolded polypeptide chains. Various studies using preformed fibrils to seed the aggregation of soluble protein molecules suggest, however, that this form of self-assembly occurs through specific intermolecular association, based on the repetition of identical or near-identical amino acid sequences^{4,5}. This specificity is consistent with the principle that ordered molecular assemblies are usually more stable than disordered ones¹, and with the idea that native-like interactions may be generally more favourable than non-native ones in protein aggregates⁶. How different, then, do two protein sequences need to be to avoid heterogeneous aggregation?

The vertebrate muscle protein titin is one of the largest known multidomain proteins, containing arrays of tandem repeats (consecutive homologous domains) of immunoglobulin domains⁷. Although titin is not associated with any known disease of protein deposition, unfolding and refolding of its domains are thought to be key aspects of its function in muscles⁸. The hypothesis that aggregation and amyloid fibril formation generally result from partially or completely unfolded proteins¹, coupled with the fact that deposited

immunoglobulin domains are associated with at least two forms of systemic amyloid disease², suggests that these immunoglobulin domains are at particular risk of forming intractable aggregates even after initial folding to their functional state is complete^{9,10}; indeed, misfolding events have been observed between adjacent identical titin domains in atomic force microscopy pulling experiments¹¹. In light of its vulnerability to aggregation in its biological environment, and because most eukaryotic proteins with multiple tandem domain repeats have a structural role³, titin provides an excellent starting point for developing an understanding of how proteins have evolved to avoid aggregation.

Incubation of the 27th immunoglobulin domain from human cardiac titin (TI I27; Supplementary Fig. S1) in a 28% solution of 2,2,2-trifluoroethanol (TFE), conditions that induce the unfolding of globular proteins while promoting intermolecular interactions¹², results in the rapid formation of structured aggregates that show characteristics of amyloid fibrils and their precursors. These aggregates have extensive β -sheet structure, as determined by far-ultraviolet circular dichroism spectroscopy, and the X-ray fibre diffraction pattern has reflections at 4.7 Å and 10 Å, consistent with a cross- β structure in which the β -strands are oriented perpendicular to the fibril axis¹³. The aggregates have a fibrillar structure by electron microscopy and bind the amyloid-specific dyes thioflavin T and Congo red, showing birefringence under cross-polarized light after incubation with Congo red.

The aggregation kinetics of wild-type TI I27, monitored by both light scattering and thioflavin T fluorescence, show a short lag phase followed by an exponential growth phase (Fig. 1a), characteristic of nucleation-dependent polymerization¹. Analysis¹⁴ of the concentration-dependent aggregation kinetics of TI I27 (Fig. 1b; see also Supplementary Fig. S2) suggests that the interaction of just two molecules is the critical step in the initial aggregation process. Using constructs with tandem repeats of wild-type TI I27, we measured the aggregation kinetics of I27–I27, I27₃ and I27₈ (2, 3 and 8 repeats, respectively)¹⁵ to investigate the aggregation of multidomain proteins. Notably, all three multimeric constructs aggregated without detectable lag phases and with elongation rates that are very similar to each other and roughly ten times faster than the aggregation rate of the monomeric construct under the same conditions (Fig. 1a). This result supports the hypothesis that the initial interaction of a pair of domains is a key step in the aggregation of this system and increases further the conundrum of how adjacent protein domains avoid aggregation *in vivo*.

The aggregation of a covalent dimer containing two non-identical immunoglobulin domains, TI I27 and its adjacent immunoglobulin domain TI I28 (I27–I28)¹⁶, was also measured (Fig. 1a). This dimer aggregates much more slowly than the I27 dimer (I27–I27), with a considerable lag phase (compare the thin black line (I27–I28) with the red line (I27–I27) in Fig. 1a). In fact, I27–I28 at 2 mg ml⁻¹, which

¹Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. ²MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

³Department of Chemistry, University of Cambridge, MRC Centre for Protein Engineering, Lensfield Road, Cambridge CB2 1EW, UK.

has a concentration of the I27 domain of 1 mg ml^{-1} , has a lag phase similar to that of the TI I27 monomer alone at 1 mg ml^{-1} , and then aggregates at a rate similar to that of TI I27 aggregating alone at 2 mg ml^{-1} . This result suggests that aggregation is initiated between I27 domains from different dimer molecules, rather than between adjacent covalently linked I27 and I28 domains, as in I27–I27. This process then allows the self-aggregation of TI I28 domains that have been brought into close proximity. Most notably, the very slow aggregation rate of I27–I28 as compared with I27–I27 suggests that the sequence identity between the domains has a major role in determining the aggregation propensity of multidomain proteins.

To investigate further the specificity of aggregation between the protein domains observed here, we devised a simple strategy to measure the extent to which non-identical sequences can coaggregate. TI I27 was incubated under aggregation conditions with equimolar concentrations of both homologous and unrelated all- β proteins that have sequence identities to TI I27 varying between 99 and 0%. The kinetics of aggregation were compared with those of wild-type TI I27 alone at the same total concentration of protein. We found that the efficiency of coaggregation increases strongly as the sequence identity to TI I27 increases (Fig. 2a–d, Supplementary Fig. S3 and Table S1). All of the proteins are unfolded in 28% aqueous TFE and form aggregates over varying timescales when incubated alone under these conditions. However, none of the domains with lower than 30% sequence identity, including TI I28, was found to coaggregate detectably with TI I27. The titin domains TI I31 and I32 partially coaggregated with TI I27, and a mutant of TI I27 aggregated with the wild-type protein as readily as it undergoes self-aggregation. The relative levels of coaggregation do not correlate with stability or with the unfolding or aggregation rates of the domains. These results therefore suggest that immunoglobulin domains with sequence identities below $\sim 30\%$ do not coaggregate detectably, whereas

those with identities above $\sim 70\%$ coaggregate as efficiently as identical domains when unfolded under these conditions (Fig. 2d).

In the cellular environment the total concentration of proteins and other macromolecules is extremely high, sometimes in excess of 350 mg ml^{-1} (ref. 17). Despite the fact that unfolding and refolding of proteins occur frequently in a biological environment, our results reveal that the diversity of protein sequences is likely to be a major factor in reducing the propensities of proteins to aggregate as the concentration of a given type of protein, except in very rare cases, is always much lower than the total protein concentration, and there is essentially no sequence identity between unrelated proteins in a given organism. Notably, the two known amyloid diseases that involve immunoglobulin domains result from circumstances in which the concentrations of the proteins involved are raised substantially above normal¹⁸. For example, AL amyloidosis is caused by the neoplastic expansion of the plasma cell population synthesizing amyloidogenic light chains, and in haemodialysis-related amyloidosis β_2 -microglobulin becomes concentrated in the post-dialysis serum of the individual, ultimately accumulating in tissue. In both disorders, the amyloid deposits have been observed to regress slowly if the amount of the precursor immunoglobulin protein returns to normal¹⁸. The potential is therefore very real for the formation of intractable aggregates by immunoglobulin domains in the biological environment.

For a protein with homologous tandem repeats, the local concentration of domains with the same fold is extremely high and hence will promote aggregation even though the total concentration of the complete protein itself may be very low. How, then, is aggregation avoided for such proteins? To address this issue, we analysed the frequency of occurrence and sequence similarities of tandem repeats in several different proteins. Tandem repeats constitute many of the proteins encoded in the genomes of organisms from all three

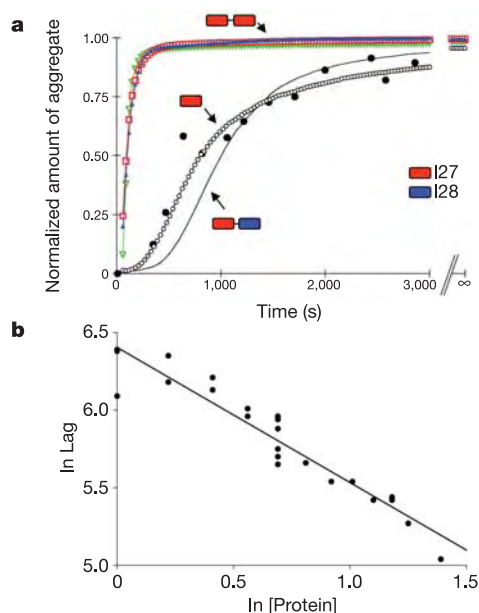


Figure 1 | Aggregation of TI I27. **a**, The aggregation kinetics of the wild-type I27 monomer monitored by light scattering (open black circles) and by thioflavin T fluorescence (filled black circles); the kinetics measured from both techniques are in close agreement. The aggregation kinetics of covalent multimeric constructs of I27 were also measured by light scattering. I27–I27, red squares; I27₃, blue triangles; and I27₈, green triangles. Aggregation of the I27–I28 dimeric construct^{15,16} is also shown (thin black line). **b**, For the monomer, the length of the lag phase scales in inverse proportion to $[\text{protein}]^{n/2}$, where n is the number of molecules involved in the critical step for initial aggregation¹⁴. The slope indicates that, for I27, $n = 2.0 \pm 0.3$ s.d..

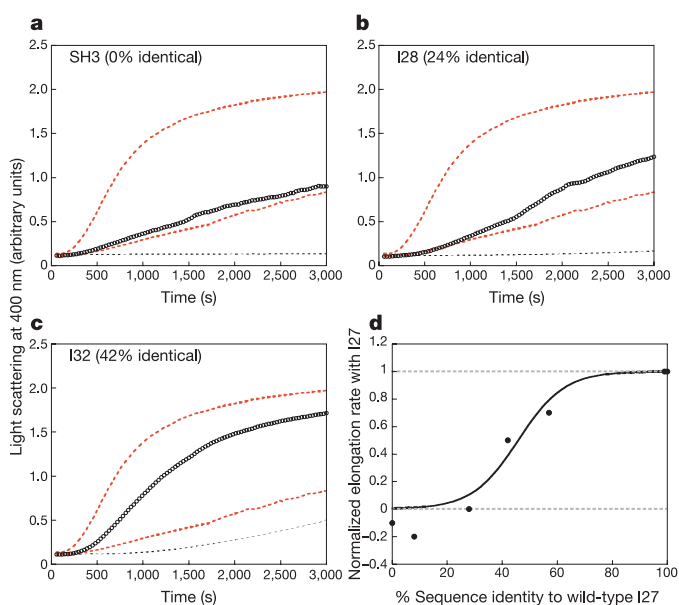


Figure 2 | Analysis of the extent of coaggregation of different proteins with TI I27. **a–c**, The aggregation kinetics (black circles) fall between two limiting rates: I27 aggregating alone at 1 mg ml^{-1} (lower broken red line) and at 2 mg ml^{-1} (upper broken red line). The aggregation time course of each protein alone at 2 mg ml^{-1} is also shown (broken black line). The protein domains used and their sequence identities (ID) to wild-type I27 were PI3-SH3, 0% ID (**a**); TI I28, 28% ID (**b**); and TI I32, 42% ID (**c**). Other data are shown in Supplementary Fig. S3. **d**, Aggregation rates from the coaggregation experiment (Supplementary Table S1) normalized and fitted to a simple sigmoidal function (unbroken line); the midpoint is at $45 \pm 3\%$ s.e.m. identity ($R^2 = 0.94$). Errors are calculated from the s.d. of repeated measurements.

kingdoms of life. Most proteins (generally more than 70% in eukaryotes) comprise more than one domain, and in vertebrate genomes about 25% of these proteins contain tandem repeats of the same superfamily (Supplementary Table S2). Moreover, in the human genome about 3% of the encoded proteins containing such repeats have arrays of ten or more repeated domains (Supplementary Table S3).

The immunoglobulin and fibronectin type III superfamilies contain some of the domains most commonly found in tandem array (Supplementary data Table S4). We therefore compared the sequence identity distributions of the domains in each of these two families, both within and between the proteins of the human genome (Fig. 3a and Supplementary Fig. S4). On average, domains within a single protein are more similar to each other than are domains in different proteins (paralogues), an indication that proteins containing several copies of a domain are likely to have arisen either from internal duplication or because domains in a protein may have similar functional constraints (for example, mechanical strength in titin I-band domains; see the sequence alignment in Table S5). Of the adjacent domain pairs in the same proteins, more than two-thirds have less than 30% identity (Fig. 3a). In addition, there is a rapid decay in the sequence identity distribution of the domain pairs such that only about 10% of the adjacent domain pairs have more than 40% identity (Fig. 3b). Particularly notable is the observation that the sequence identity between adjacent domain pairs is significantly

lower than the sequence identity between non-adjacent domains in the same protein, suggesting that adjacent domains are under greater evolutionary pressure to diverge than non-adjacent domains (Fig. 3a). Our analysis of the domains of proteins in the human genome in general, and of the titin domains in particular, provide an overview of the sequence dependence of aggregation. As only part of any protein molecule is incorporated in the structured regions of an aggregate, however, some residues will be more important than others in promoting or inhibiting aggregation¹, and overall sequence variations will only be averages of local variations. This fact means that individual cases are expected to have characteristics that are dependent on local factors in addition to the generic ones that we discuss in this work.

Our experiments therefore show that, although immunoglobulin domains with closely similar sequences have a high tendency to coaggregate, when the sequence identity falls below a threshold of 30–40% the propensity to coaggregate is negligible. Although other factors are undoubtedly also important, this finding can explain why a single dominant protein is almost invariably found in proteinaceous deposits formed *in vivo*¹⁹, including amyloid fibrils and plaques, and why the ability of a given protein to promote amyloid formation by a different type of protein through seeding is strongly dependent on the similarity of their sequences⁵. These findings also indicate that intractable aggregates, such as amyloid fibrils, are stabilized by highly specific interactions in addition to the dominant main chain interactions that are common to all polypeptides. Such specificity is reflected by the well-defined conformations and close-packed nature of the amino acid side chains in amyloid structures^{20–22}, indicating that one particular packing out of the many alternatives consistent with a cross- β structure is preferentially adopted under given conditions. As with the more familiar three-dimensional crystals, the most stable structure of a one-dimensional fibrillar aggregate is very likely to involve highly regular repetitive arrays of identical molecules.

Although the sequence identity between adjacent domains in multidomain proteins is almost invariably less than 40%, there are some domains for which the identity is high enough for coaggregation to become a potential problem. For some of these proteins there may be other sequence-dependent effects that inhibit aggregation: such mechanisms include the conservation of specific residues that act specifically to prevent aggregation^{23–25}, the incorporation of disulphide bonds that add constraints to the adoption of non-native structures¹, and the ability to fold sequentially²⁶. More generally, other mechanisms that are present in living systems to prevent aggregation, including molecular chaperones²⁷ and the unfolded protein response²⁸, will deal with such situations. (These mechanisms are likely to be of particular importance during protein biosynthesis where there may be relatively high local concentrations of identical proteins.) Our results, however, suggest that such mechanisms will normally be required to deal primarily with those situations in which the intrinsic sequence-based mechanisms are insufficient to prevent aggregation. For multidomain proteins, for example, a likely situation is where gene duplication has occurred relatively recently in evolutionary history. In this context, the role of molecular chaperones in shielding aggregation-prone mutational variants prior to compensating mutations has been shown to be an important aspect of the evolution of proteins with new functionalities²⁹.

In summary, our findings suggest that maintaining a low sequence identity between proteins is an important evolutionary characteristic that strongly inhibits misfolding and aggregation in the crowded environment of a living cell. In the specific case of the multidomain proteins that dominate the genomes of higher organisms, bioinformatics analysis shows that most adjacent domains are from different fold families and therefore have sequence identities well below the threshold for significant coaggregation. For tandem repeats of a given domain, many of which are immunoglobulin-like domains such as those studied here, the local effective protein concentration is high

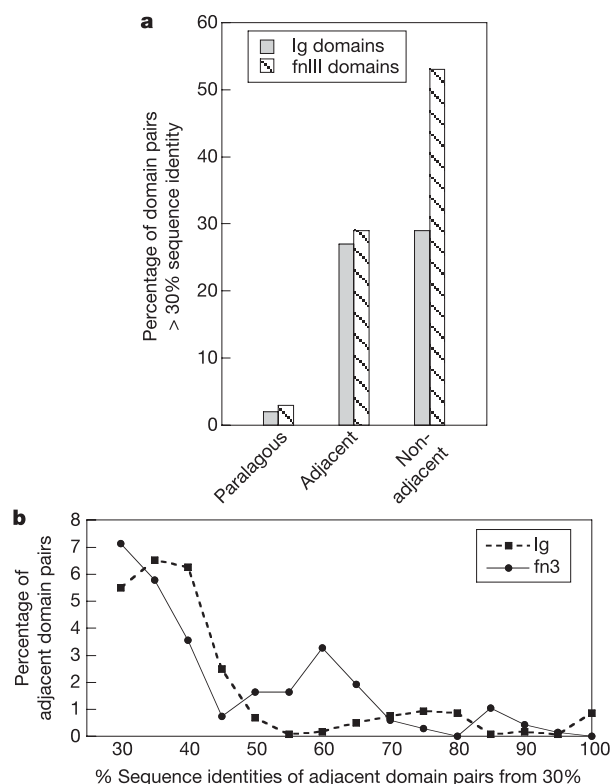


Figure 3 | Analysis of sequence identities of immunoglobulin and fibronectin type III domains in the human genome. Domains that are in different proteins in the same organism are termed 'paralogous'. Domains in the same protein are 'adjacent' if positioned within 30 residues of each other and 'non-adjacent' if positioned more than 30 residues apart. **a**, Percentage of domain pairs with more than 30% sequence identity in the immunoglobulin (Ig) and fibronectin type III (fnIII) superfamilies. Most of these domains in separate proteins in humans have less than 30% sequence identity to each other. **b**, Sequence identity distributions of adjacent domain pairs in the Ig (broken line) and fnIII (unbroken line) superfamilies. Only ~10% of adjacent Ig and fnIII domains have more than 40% sequence identity. Full details of the alignment analysis are given in the Supplementary Information.

and coaggregation becomes a particular risk. Our results, however, suggest that, in addition to the evolutionary pressure for a protein to conserve specific positions of its sequence for function, folding and structural stability, the diversification of protein sequences to avoid unwanted interdomain or intermolecular association could be an important factor in molecular evolution.

METHODS

Aggregation and light scattering. All aggregation experiments were carried out in 28% TFE in PBS buffer (pH 7.4) at 25 °C. Protein was purified as described^{15,16}. The aggregation kinetics of the wild-type TI I27 monomer were monitored by light scattering using a Synergy Bio-Tek plate reader and by thioflavin T fluorescence using an Aminco Bowman fluorimeter. All solutions were filtered immediately before use. Full details of the kinetic analysis are given in the Supplementary Information.

The aggregation kinetics of covalent multimeric constructs of TI I27 were also measured by light scattering. All solutions contained 2 mg ml⁻¹ of protein (that is, equimolar concentrations of the titin domains). The domains in these multimeric constructs are each linked by 2–4 amino acids, which are required for DNA ligation¹⁵. In the coaggregation experiments, each monomeric protein was added to a solution of TI I27 to reach a final concentration of each domain of 1 mg ml⁻¹ (a total protein concentration of 2 mg ml⁻¹). The extent of light scattering was monitored as a function of the time of incubation.

Sequence analysis. For the sequence analysis, we used the structural domain assignments to complete genome sequences in the SUPERFAMILY database, version 1.61 (ref. 30). Full details are given in the Supplementary Information.

Received 14 March; accepted 1 September 2005.

- Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
- Selkoe, D. J. Folding proteins in fatal ways. *Nature* **426**, 900–904 (2003).
- Apic, G., Gough, J. & Teichmann, S. A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325 (2001).
- Biere, A. L. et al. Parkinson's disease-associated α -synuclein is more fibrillogenic than β - and γ -synuclein and cannot cross-seed its homologs. *J. Biol. Chem.* **275**, 34574–34579 (2000).
- Krebs, M. R. H., Morozova-Roche, L. A., Daniel, K., Robinson, C. V. & Dobson, C. M. Observation of sequence specificity in the seeding of protein amyloid fibrils. *Protein Sci.* **13**, 1933–1938 (2004).
- Silow, M., Tan, Y.-J., Fersht, A. R. & Oliveberg, M. Formation of short-lived protein aggregates directly from the coil in two-state folding. *Biochemistry* **38**, 13006–13012 (1999).
- Labeit, S. et al. A regular pattern of two types of 100-residue motif in the sequence of titin. *Nature* **345**, 273–276 (1990).
- Linke, W. A., Stockmeier, M. R., Ivemeyer, M., Hosser, H. & Mundel, P. Characterizing titin's I-band Ig domain region as an entropic spring. *J. Cell Sci.* **111**, 1567–1574 (1998).
- Raffen, R. et al. Physicochemical consequences of amino acid variations that contribute to fibril formation by immunoglobulin light chains. *Protein Sci.* **8**, 509–517 (1999).
- McParland, V. J. et al. Partially unfolded states of β_2 -microglobulin and amyloid formation *in vitro*. *Biochemistry* **39**, 8735–8746 (2000).
- Oberhauser, A. F., Marszalek, P. E., Carrión-Vasquez, M. & Fernandez, A. Single protein misfolding events captured by atomic force microscopy. *Nature Struct. Biol.* **6**, 1025–1028 (1999).
- Chiti, F. et al. Designing conditions for *in vitro* formation of amyloid protofibrils and fibrils. *Proc. Natl Acad. Sci. USA* **96**, 3590–3594 (1999).
- Sunde, M. et al. Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* **273**, 729–739 (1997).
- Oosawa, F. & Asakura, S. *Thermodynamics of the Polymerization of Protein* (Academic, London, 1975).
- Steward, A., Toca-Herrera, J. L. & Clarke, J. Versatile cloning system for construction of multimeric proteins for use in atomic force microscopy. *Protein Sci.* **11**, 2179–2183 (2002).
- Scott, K. A., Steward, A., Fowler, S. B. & Clarke, J. Titin: a multidomain protein that behaves as the sum of its parts. *J. Mol. Biol.* **315**, 819–829 (2002).
- Ellis, R. J. & Minton, A. P. Join the crowd. *Nature* **425**, 27–28 (2003).
- Pepys, M. B. Pathogenesis, diagnosis and treatment of systemic amyloidosis. *Phil. Trans. R. Soc. Lond. B* **356**, 203–211 (2001).
- Rajan, R. S., Illing, M. E., Bence, N. F. & Kopito, R. Specificity in intracellular protein aggregation and inclusion body formation. *Proc. Natl Acad. Sci. USA* **98**, 13060–13065 (2001).
- Nelson, R. et al. Structure of the cross- β spine of amyloid-like fibrils. *Nature* **435**, 773–778 (2005).
- Jaroniec, C. P. et al. High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. *Proc. Natl Acad. Sci. USA* **101**, 711–716 (2004).
- Krishnan, R. & Lindquist, S. L. Structural insights into a yeast prion illuminate nucleation and strain diversity. *Nature* **435**, 765–772 (2005).
- Otzen, D. E., Kristensen, O. & Oliveberg, M. Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc. Natl Acad. Sci. USA* **97**, 9907–9912 (2000).
- Steward, A., Adhya, S. & Clarke, J. Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *J. Mol. Biol.* **318**, 935–940 (2002).
- Richardson, J. S. & Richardson, D. C. Natural β -sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99**, 2754–2759 (2002).
- Netzer, W. J. & Hartl, F. U. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388**, 343–349 (1997).
- Barral, J. M., Broadley, S. A., Schaffar, G. F. & Hartl, F. U. Roles of molecular chaperones in protein misfolding diseases. *Semin. Cell. Dev. Biol.* **15**, 17–29 (2004).
- Goldberg, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895–899 (2003).
- Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).
- Gough, J. & Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30**, 268–272 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Scott, A. Steward and J. Zurdo for clones; and R. Bader, D. Hall, C. Ponting and C. Chothia for discussions. This work was supported by the Wellcome Trust and the Leverhulme Trust (C.M.D.). C.F.W. held a Wellcome Trust prize studentship and J.C. is a Senior Wellcome Trust research fellow.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.C. (jc162@cam.ac.uk) or C.M.D. (cmd44@cam.ac.uk).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Seeking soft skills

Scientists who want to pursue a career outside academia need a 'PhD plus', panellists told a career seminar last week at the European Biotech Crossroads conference in Lille, France. The 'plus' is any skill beyond the scientific, from project management to writing. But where do you get that plus?

The panellists' own experiences illustrated different avenues — although all of them involved little or no formal training in these off-the-bench skills.

Manuel Gea, chief executive of Bio-Modeling Systems in Paris, learned about biotech and information technology by interacting with colleagues at other jobs. Tristan Rousselle, president of Protein'eXpert, a biotech company in Grenoble, France, had scientific skills but no management training when he launched his company, so he learned along the way. And Terry Vrijenhoek, a PhD student at Radboud University's medical centre in Nijmegen, the Netherlands, learned organizational skills, writing and graphic design when he helped to set up the Genomics Network for Young Scientists.

But nowadays you don't have to take the do-it-yourself

route. There are a lot of formal courses, programmes and additional degrees that will teach you off-the-bench skills — many of which we have covered in our Postdocs & Students features (www.nature.com/naturejobs/archive/archive-pdns.html). But scientists at every stage of their career would do well to think about the additional skills they use to get their science done. Excelling in these can help you to climb the tenure track, but they also transfer well to other sectors.

Although formal training is great, it isn't essential. People looking to hire scientists outside academia want someone who can take on responsibilities, said Rousselle, and they don't necessarily care how the person got the skills needed to do the job. They will also respect people who have taken the initiative to gain those additional skills by whatever means.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

US Head Office, New York
 345 Park Avenue South, 10th Floor,
 New York, NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager/Corporations:
 Peter Bless
 Classified Sales Representatives
 Tel: +1 800 989 7718

**New York/Pennsylvania/
 Latin America:** Andrew Bennie
**Midwest USA/Maryland/
 NIH:** Wade Tucker
East USA/Canada:
 Janine Taormina

San Francisco Office
Classified Sales Representative:
 Michaela Bjorkman
 225 Bush Street, Suite 1453
 San Francisco, CA 94104
 Tel: +1 415 781 3803
 Fax: +1 415 781 3805
 e-mail: m.bjorkman@naturesf.com

European Head Office, London
 The Macmillan Building,
 4 Crinan Street,
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director: Nevin Bayoumi (4978)
European Sales Manager: Andy Douglas (4975)

Advertising Production Manager: Billie Franklin
 To send materials use London address above.
 Tel: +44 (0) 20 7843 4814
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
 Patrick Phelan
 e-mail: p.phelan@nature.com

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

After eight hours in the lab, molecular biologist Joaquin Espinosa goes to his other job — a 12-hour night shift. Granted, he is allowed to sleep some of the time, but it is no less demanding than his day job, as he is taking care of his newborn twin daughters, Paloma and Azul.

Espinosa directs his fledgling lab at the University of Colorado, Boulder, from noon until 8 at night. Then he heads home to help wife Eliana Gomez bathe, change and feed their girls, taking responsibility until 8 in the morning, when his mother takes over. The schedule works better for Espinosa than full-time paternity leave.

"I'm wasted in the mornings, so I sleep in and get to the lab around noon," he says. "This is one of the privileges of the academic setting." Indeed, the flexibility of academia can sometimes offset the pressures of balancing home and lab time.

All families — large or small, with newborns or adolescents — must have some time together. Lab-management experts have found ways of keeping that time a priority during the hectic years before gaining tenure. They also manage to squeeze some work in at home without crossing the workaholic line. And some have invented creative ways to harmonize home and lab by involving family.

Time for tots

Juggling a new career and starting a family worries many scientists. But waiting to start a family until after tenure is a reproductive gamble.

"It takes nine months to go on the job market and, coincidentally, it takes about nine months to make a baby," says Sandra Schmid, chair of cell biology at the Scripps Research Institute in La Jolla, California. She had her son Jeremy, now 18, at the end of her postdoc and daughter Katie, now 14, as a new investigator.

Espinosa and Gomez took a slightly different tack. Gomez decided to take a year off after her postdoctoral fellowship and to look for a scientific position next year. "It would be way too much work setting up two labs and bringing up babies at the same time," says Espinosa. The approach became essential once they learnt they were expecting twins.

Schmid advises parents of toddlers to rely as much as possible on others, including students, postdocs, spouses, relatives, technicians, administrative help and hired help at home if necessary. "You don't have to do it all on your own," she says — the value of your time has increased since graduate school.

"Don't hesitate to say, 'Ask me to join that committee again in two

Toddlers, teens and test tubes

Young careers and young families can leave investigators feeling stretched.

Kendall Powell finds out how to keep the two from clashing.

years when Billy starts kindergarten,'" she adds, noting that it gets easier as children go to school and can help in the house. And count your blessings when it comes to schedules — Schmid coached her kids' soccer teams for seven years because she could leave work at 4 p.m. every Monday.

Coaching or taking a day off to chaperone a school field trip may mean more to children than staying at home full-time. Although it may take some ingenuity, busy researchers don't have to sacrifice quality time. Find out what matters most to your children and then make time for swimming lessons or tea parties.

Drawing the line

Get over the guilty-at-home, guilty-at-work syndrome and be focused on where you are, advises Schmid. Susanne Mandrup, a molecular biologist at the University of Southern Denmark in Odense, agrees. In the early years, she says, "you often find that home gets the lowest priority. You have to manage that, otherwise you ruin your personal life."

When at home, Mandrup gives her family her full attention, limiting television and phone calls and finding activities that her teenagers can enjoy with her and her husband. "We concentrate on each other, do daily things such as cooking dinner or baking bread together," she says. Her family make time together each night as well, in the Danish tradition of cosy intimacy called *hygge*. Before the youngest child goes to bed, this might include reading, watching a film or playing a game accompanied by some tea and chocolate.

Carol Thornber, a marine ecologist at the University of Rhode Island in Kingston, reserves weekend time for her husband who, during the week, works three hours away in Amherst, Massachusetts. Designating family hours keeps her and other junior scientists motivated to be efficient at work.

"You have to know when to listen to your wife's 'no laptop on holiday' policy," notes Dirk Schübeler, an epigeneticist at the Friedrich Miescher Institute in Basel, Switzerland. But when his young sons go to bed, he catches up on e-mail with his US collaborators.

Home office

Parents of young children often have to work from home — Espinosa and Mandrup spend the hours after their children's bedtime reading papers or writing. "For many years, I've worked every night to catch up on what I don't get done during the day," says Mandrup.

Espinosa changed the way he uses his computer so that he can work as easily in his home office as in his lab office. "Before, I didn't even bring my computer home



Starting young: Madeline Hull Brazas gets to work.

C. HULL

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

and I would unplug my brain from science on Friday evening. But I can't afford to do that anymore," he says.

He set up a home studio and switched to the smallest laptop available. He equipped both offices with large monitors, extra keyboards and hard drives to back up files. This helps him keep in touch with lab members, too, whose schedules might not overlap much with his.

Working at night or in the early morning while family members sleep can reduce the impact of work on quality time — find a schedule that works for you and your family. Schübeler prefers starting and leaving work early, in contrast to Espinosa's late start at the lab.

Christina Hull, a molecular biologist at the University of Wisconsin, Madison, routinely goes in extra early on weekend days to check experiments. This saves the rest of the day for family activities with her husband and six-year-old daughter Madeline.



In the pink: late nights don't faze Susanne Mandrup.

"Last Sunday I got in at 5 a.m. and had everything done by 7 a.m., in time to make breakfast with friends," she notes. As her daughter gets older, Hull tries to incorporate necessary weekend trips to the lab with some fun for Madeline.

Daughter-at-lab day

Some faculty members would prefer to separate home life from work with a clearly defined line. But others, including Hull, say that there are ways to bring your two halves — scientist and parent — together.

Madeline, for instance, has her own box of 'experiments' to investigate while waiting for her mother to finish bench work. She gets her own agar plates, gloves and different coloured solutions in test tubes to test the growth of common bacteria, moulds and mildews.

"Now I have a six-year-old who absolutely loves to come to the lab," says Hull. "Making it a fun environment and including her has helped."

"It's important for your kids to see that their parents enjoy work and find it exciting — it's an important impression to convey," says Mandrup. It also helps them accept your long working hours.

Your family might even accept work meetings or phone calls at home, if it lets you spend more time with them, says Bernard Golding, an organic chemist at the University of Newcastle upon Tyne, UK. He often meets colleagues at home because his wife is "more agreeable to this than my staying late".

Golding also found an inventive way to expose his 11-year-old granddaughter Holly to his work: she went with him and his wife to a chemical-biology meeting at the Austrian ski resort of Kleinwalsertal. "She got on really well with five of my trainees there. It was a good message for her to mix with these academically minded, clever, attractive personalities," he says.

Golding wishes that he had taken his children to more meetings and he now encourages junior faculty members to look for meetings with open times during the day as a "good way of involving family in your job". He also advises parents not to cut out exercise when family and work duties expand. "If the going gets tough, at least try to walk or cycle to work," he says. Combining chores with leisure can help: Mandrup makes dinner while discussing her teens' days, and Espinosa reads papers while waiting for the midnight feed.

"Nobody ever talks about, 'Gee, I had a balanced week,'" says Schmid. "The meaningful phrase is a balanced life. You can't do everything all the time at all stages. Look overall at what you can do with a young family and career."

Kendall Powell is a freelance science writer based in Broomfield, Colorado.

Transport of delight

The highs and lows of commuting.

Roland Denison

The cyclist flies downhill towards the city in a semi-trance of endorphins and adrenaline. Traffic lights change a fraction before he would have reached the brakes as he shoots past hospital, shops and park. From here the road runs flat and straight, aimed at the heart of the old city. He slips his derailleur down and rises from the saddle to maintain speed. A week ago he couldn't have done that. Even yesterday he would have been navigating a fractured route between the bumpers of stationary vehicles; an inconstant path whose gaps and alleys closed and opened without warning.

But today is special. Today there is no traffic. The roads belong to him alone.

Moving the citizenry around had never been simple. Centuries before, carts had been forbidden entry to the city during the hours of daylight. The bridges over the river, whose oily tang he can already smell, bore toll-gates. In the current century that idea was reinvented. Batteries of cameras standing guard, noting down entries in their electronic ledgers and tallying up their fee. Soon every road in the metropolitan area, indeed the whole country, had its own tariff. To keep track of this revenue, vehicles were fitted with radio tags to pinpoint their position to within half a metre. And still the jams worsened, average speeds dropped, and drivers seethed.

The cyclist raises his cadence for his favourite part of the journey: two inter-linked gyratories, once the boundary of the original levy. Slowing slightly, he leans into the initial left-hand bend, right foot down to avoid clipping the tarmac. Then to the right, straightens for eight cranks to maintain momentum for the second left-right-left combination. Ten seconds of bliss that any bobsleigh pilot would envy.

Half a mile on, a cluster of side streets form the shortest path to the river. Here the presence of pavements makes corners difficult to judge. He can just remember when all roads in the city had pedestrian-thronged sidewalks, but their users were

long since driven away by the dangers of the adjacent road. Then the roads were widened to engulf the redundant walkways. But, as so often before, larger roads encouraged more traffic and greater congestion.

Drivers had felt discriminated against by the tagging and tolls, their human rights infringed. The judiciary agreed, but rather than abolish the offence, the government abolished the discrimination, extending it to the entire populace. Personal identity cards became subcutaneous identichips. An entire population located to within

have, if they dare try, a glorious life." He has read this every day for years but it was only last week that he had grasped its meaning. Yet a glorious life seemed too ambitious: but a glorious day, he could dare that much.

The CTC had long since calculated a solution to the city's traffic woes. Traffic created congestion, vehicles created traffic — banish the vehicles, abolish the jams. But that wasn't politically acceptable. Instead, the data logs on which the CTC based its calculations had to be amended. A lie of omission to keep the CTC labouring in the vain hope that the traffic could be made to flow. That was the cyclist's job.

Standing on the pedals he ascends the bridge crossing the wide river in glorious isolation before plunging between the artificial cliffs on the north bank. Past theatres, banks, museums, bars and shops, his progress continuous and unimpeded. Today the streets are his — for yesterday he had dared the truth. He had passed the logs unexpurgated to the CTC and it had drawn its inevitable conclusion.

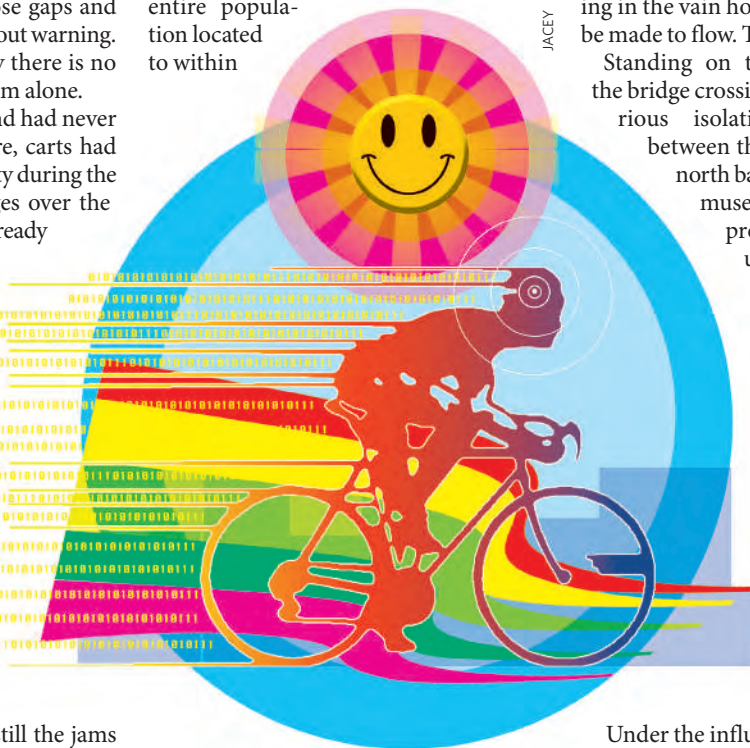
He nears his office overlooking the canal, a previous century's attempt at traffic management. He dismounts on the move, jogging a few steps to bring the bicycle to a stop.

Under the influence of the enlightened CTC, the periphery of the city is currently populated with artificially contented drivers travelling in futile circles. Today the city streets are the cyclist's domain, and it feels good.

As he reaches the door the euphoria that has carried him through his journey evaporates, replaced by a weight of dread. He turns, remounts his bike and cycles back into the city, his mood lifting immediately with a fresh endorphin release.

Today he belongs to the roads and their call cannot be denied.

Roland Denison has recently left the XVth Directorate of the European Commission for the Wageningen offices of the United Nations. His novels featuring the alternative-realities private detective Giles Coppice continue to attract a cult following.



JACEY

20 centimetres, day and night.

All the information was processed by a central traffic computer, the CTC, which the cyclist was paid to deceive. At first it simply advised in-vehicle navigation systems, but that had little effect. Everyone had a better idea on how to beat the jams. Everyone knew a secret shortcut. No one followed its advice. The G5 riots changed that. Hastily passed legislation upgraded citizens' identichips to release endorphins and their antagonists under remote control. Now the CTC could make drivers feel good about following its route planning and depressed for ignoring it, but even so the traffic didn't move.

The cyclist passes a motto, written on the side of a 300-year-old church: "All may